Improved Galaxy Counting Techniques and Noise Reduction Algorithms as Applied to the Sloan Digital Sky Survey

by

Michael A. Specian

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2015

© Michael A. Specian 2015

All rights reserved

Abstract

In the last two decades, the amount of information gathered by galactic surveys has increased dramatically. As statistical uncertainty declines, systematic effects become the primary impediment to achieving precision cosmology. In this dissertation, we examine three classes of problems that introduce percent-level systematic noise in the galaxy clustering catalogs of the Sloan Digital Sky Survey (SDSS). First, we uncover numerous errors in the photometric and spectroscopic footprint definitions of the SDSS's 6^{th} data release. We correct these errors, producing the most accurate description of the SDSS's survey geometry ever. Next, we address the problem of counting galaxies in cells when those galaxies lack spectroscopic redshifts. We test a variety of counting techniques across a range of cell sizes, redshifts, and regions in order to recommend a set of best counting practices. Finally, we introduce a novel noise cleansing algorithm that uses a Bayesian methodology to estimate the most likely values of statistical and systematic noise given noisy data and best-guess fiducial signal and noise models. We conclude by combining these three solutions to estimate the true signal of real galaxy data drawn from the Sloan Digital Sky Survey.

ABSTRACT

Reader 1: Professor Alex Szalay (Advisor), Physics & Astronomy

Reader 2: Professor Yanif Ahmad, Computer Science

Reader 3: Professor Charles Meneveau, Mechanical Engineering

Reader 4: Dr. Marc Postman, STScI

Reader 5: Professor Mark Robbins, Physics & Astronomy

Acknowledgments

The study of physics is an investigation into how a set of initial conditions produces a final result. If this doctoral dissertation is my final result, then I want acknowledge and thank those that helped make it happen.

My "initial conditions" during childhood made it unlikely that I would ever reach this point. Yet I was exceptionally lucky to grow up in the beautiful little borough of Alpha, NJ. Alpha Public School provided me an excellent elementary education, but what really saved me was the Alpha Public Library and its librarian, Myrna Minardi. Myrna offered me refuge and support as if I was her own grandson. I am not sure I could have made it through without her. She is a testament to the value of small town public libraries everywhere.

My former high school US History teacher Laurie Schmid has benefited my life in more ways than I can count. She welcomed me into her family, provided me a place to live, and helped pay for my final semester of college when I ran out of money. She discovered and cultivated my potential as a debater, an activity that has changed my life. My accomplishments are her accomplishments.

I would not be where I am today without the support of my undergraduate thesis advisor

ACKNOWLEDGMENTS

Tereasa Brainerd. Living in Boston during the summer while also trying to earn enough to pay the next year's tuition was daunting. Yet Professor Brainerd mustered the generous financial support to make my summer research possible. Whatever career I have a scientist going forward, I owe to her.

No one has been more instrumental to my progress through graduate school than Alex Szalay. I consider myself extremely lucky to have been able to study alongside someone of his intellectual caliber. Through the years, Alex has offered unwavering financial support. He assembled a team and built a computational infrastructure without which this research project does not occur. While many advisors funnel their students in a direction they see fit, Alex always encouraged me to follow my non-cosmological interests, like science policy. Few students are as fortunate to receive the level of academic freedom that Alex provided to me.

The team Alex assembled at JHU and his colleagues around the world all deserve recognition in helping me complete my research. I want to thank Anthony Banday for tolerating my summer abroad, and for helping me gain valuable insight about my career as a scientist. After hitting MANY theoretical brick walls (see Appendix G), Don Geman donated his time and finally helped me turn the corner. For that I am very grateful. Numerous other graduate students and postdocs provided advice and support along the way. I particularly want to recognize Tamas Budavári, Manuchehr Taghizadeh-Popp, Mark Neyrinck, Domenico Tocchini-Valentini, and all the IDIES nerds who kept the computer systems up and running. I would also like to thank the PHA staff, who work quietly behind the scenes

and make everything possible. Their efforts are not recognized nearly enough.

However, no one has provided me with more emotional support throughout this dissertation process than Chitra Kalyandurg. The love and support she provided me are unmatched. Her selflessness and generosity kept me from going crazy, and I regret that her efforts were not reciprocated nearly as much as she deserved. This dissertation is a product of her sacrifice as much as mine. I love you, Chitra.

Finally, thank you to all other friends and family who provided support and encouragement over the years especially Christiana Jackson (better late than never), Kait Houlihan, Lynn Carlson, and Jacob Koskimaki. You guys are all right!

Dedication

This dissertation is dedicated to all the students, scientists, funding agencies, and public supporters that power our scientific enterprise. You make the world go round.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The origin and meaning of the night sky has inspired no shortage of legends, myths, and theories. Twinkling points of light, scarce to be counted, rotate overhead with reliable order. Configurations of constellations and asterisms are recognized as avatars of gods and heroes. Each planet predictably follows its course, occasionally in retrograde, forming configurations with predictive power. Comets arrive periodically, often as omens of death or catastrophe.

Throughout most of human history, interpreting the nature of the cosmos lay squarely in the jurisdiction of philosophers and religious authorities. While science and mathematics had demonstrated by the 3^{rd} century BC that the Earth was spherical, the nature of its place in the Universe remained largely unknown for the next two millennia.

In the early 17^{th} century, the telescope was invented and purposed to identifying ships as they crossed the horizon towards port. Italian philosopher and mathematician Galileo

Galilei took that invention and pointed it towards the sky. He gathered a small data set, discovering that the moon's surface, rather than being a uniform sphere, was filled with craters, mountains, and other nonhomogeneous features. In observing solar sunspots, he dispelled the notion of the sun as a perfect spherical body. He found four tiny dots of light circling Jupiter, objects that we now identify as the Jovian moons Ganymede, Callisto, Europa and Io.

Galileo's observations that celestial objects were "imperfect" and could orbit something other than the Earth sparked a major paradigm shift — one that recognized the Sun, not the Earth, as the center of the Universe. This data provided critical corroborative evidence for the heliocentric theory of 16^{th} century mathematician and theoretical astronomer Nicolaus Copernicus. In tandem, theory and data had begun to expose the reality of the Universe.

In Prague, a German mathematician named Johannes Kepler had come to possess the scientific logs of Danish nobleman and astronomer Tycho Brahe. In the 1570's, Brahe had begun an ambitious plan to manually map the night sky with unprecedented accuracy. At the time, the prevailing model of celestial motion was that of Ptolemaic epicycles, which held that the Sun and planets traveled on a series of nested circular orbits around Earth (an idea we recognize today as the expansion of a two-dimensional function into a sum of Fourier modes).

Kepler was greatly influenced by the discoveries of Galileo. Those discoveries provided a new context within which to interpret Brahe's observations. Kepler subsequently proposed that planets traveled in elliptical orbits around the Sun, leading to the development

of what we now refer to as "Kepler's laws of planetary motion." Isaac Newton would soon go on to derive these laws from first principles using his revolutionary theory of gravitation.

The combined efforts of Copernicus, Galileo, Kepler, and Newton reorganized the entire framework of astronomy into a new paradigm (the likes of which are described in Thomas Kuhn's seminal work, *The Structure of Scientific Revolutions* (Kuhn, 1996)), one from within which all future considerations of the topic would occur. It also marked the birth of astrophysics as a data-driven science.

Over the next 300 years, technological limitations meant that progress in astrophysics was necessarily slow. Unlike other scientific fields like biology, chemistry, and physics, the vast majority of astronomical experiments could not be conducted in laboratories. Astronomers operated under the constraint of having almost all of their data be accessible only the form of light. And even then, only two broad ranges of radiation — optical and radio — permeated Earth's atmosphere.

By the dawn of the 20^{th} century, telescopic design had matured beyond refracting lenses, which bore the unfortunate burden of deforming under their own weight. Large, smooth, finely-curved mirrors to serve in reflecting telescopes, like those at the Mount Palomar Observatory in California, were constructed. Light was split into its constituent wavelengths using diffraction gratings (due in no small part to the early efforts of Johns Hopkins's Henry Rowland) and spectrographs. Emission lines were identified that served as signatures of physical processes, such as the HI 21 cm line that marked the presence of neutral hydrogen.

These advancements in engineering allowed scientists like Edwin Hubble, Fritz Zwicky, and Penzias & Wilson to reach breakthrough conclusions. We learned that the Universe was immensely bigger than anyone had imagined upon discovering a multitude of galaxies beyond our own Milky Way. Using the Doppler effect, astronomers deduced that galaxies were speeding away at rates roughly proportional to their distances. This prompted the realization that rather than being a static entity, the Universe was expanding. The subsequent discovery of a nearly uniform "background" of microwave radiation lent support to the Big Bang theory and estimates that the Universe was in fact billions of years old.

Today we have the ability to place telescopes in orbit above the Earth's atmosphere. Studies of supernovae observed by the Hubble Space Telescope revealed that not only was the Universe expanding, but that expansion was accelerating. The energy behind the acceleration came be known as "dark energy" and its discovery signaled that there was yet another component of the Universe to account for.

Questions began to accumulate: How did the Universe begin? What is its shape and why does it appear to be so flat? Were primordial density fluctuations Gaussian in nature? What were their amplitudes? Was the early Universe "tilted" in one way or another? How fast is it expanding? How did the densities of matter and radiation change over time? What was the role of neutrinos? And just how did we arrive at the distribution of galaxies that we observe today?

The answers to these questions were encoded in values that would come to be known as "cosmological parameters". The only way to accurately quantify these parameters was

through the acquisition of data — and massive amounts of it.

Twenty-five years ago, due to a paucity of data, astrophysicists would have counted themselves lucky to achieve a 50% error bar on a measurement. But just as it had done before, the field of astronomy was about to experience a paradigm shift accomplished in no small part by scientists involved in the Sloan Digital Sky Survey (SDSS).

The SDSS project was charged with mapping one quarter of the entire sky from its perch in Apache Point, New Mexico. During the eight years of its primary observing run Sloan would gather about 1 million spectra, or approximately 10 times more than had been gathered in all of history to that point. Other surveys like the 2dF Galaxy Redshift Survey and DEEP2 would add to that number.

This explosion of data induced a paradigm shift in computational analysis that has, in various contexts, been referred to as Big Data, eScience, data-intensive discovery, and the Fourth Paradigm. No longer constrained by a dearth of data, scientists were now inundated by more than they could process. This prompted the development of numerous algorithms and database technologies that would revolutionize the field. Finally, astronomy was becoming a true precision science.

1.1 Approaching Precision

The SDSS and associated computational tools have been a boon for astrophysics. Yet as with all scientific fields, capacity for short and medium-term improvement is not unlimited.

Technological innovations are inhibited in large part by funding constraints. Tycho Brahe was able to finance his work because he possessed at one point (it is estimated) one-tenth of all the wealth in Denmark. American scientists are largely beholden to funding agencies like the National Science Foundation, NASA, and the Department of Energy. Despite legislative promises to double federal allocations for scientific research, the financial collapse of 2008, federal budget deficit, and general political sclerosis have caused science funding to stagnate while many promising research initiatives go unrealized.

Under these circumstances, it is important to develop sophisticated statistical tools to extract maximum information from the data we do possess. Within this dissertation we attack the question of precision science from three distinct perspectives — galaxy redshift survey footprint corrections, mapping galaxies in three dimensions when their radial depths are unknown, and statistical and systematic noise reduction.

We limit the majority of our analysis to the SDSS's sixth data release (DR6). We choose this particular survey release for a few reasons. First, the survey is decidedly complete and unlikely to experience many future adjustments. The geometry of the survey is provided in a way that allows its photometric and spectroscopic footprints to be studied in detail.

DR6 was the sixth of eight official data releases for the original survey. At the time, it had progressed to a state in which it had probed the majority of its ultimate volume. Redshifts were known for hundreds of thousands of galaxies, enough to draw meaningful statistical conclusions. By falling in the midst of other data releases, the survey was dynamic. It was inevitable that new data would be added to that which already existed. Given

that the most fruitful period for surveys is likely to be while they are still active, it is important to note how intra-release assumptions and observational protocols can leave errors behind.

Much of the SDSS geometry background is provided in Chapter 2. As we uncover in Chapter 5, its footprint contains a significant number of mistakes, most of which are being revealed for the first time within this dissertation. These mistakes can be characterized into 3 classes — regions that should have been included but weren't, regions that should have been excluded, and unreported sampling anisotropies. It is our hope that their discovery and resolution can offer object lessons in improving survey design, or at least the ways in which data releases are presented to the greater scientific community.

Next, in Chapters 3, 4, and 7, we address the issue of uncertainties. In general, there are two types of uncertainties that limit our ability to achieve arbitrarily high-precision results — statistical and systematic. Systematic errors are biases that result when repeated measurements of a given quantity differ from the mean, such as when an instrument is imperfectly calibrated. For many applications in the history of astronomy, statistical errors, those characterized by a variance of a measured quantity about its mean, have been of greater concern.

As surveys have become larger and more robust, statistical noise has declined, often leaving systematic effects as the largest remaining sources of uncertainty.[1] To extract the optimal amount of useful information from future measurements it is imperative, therefore,

[1]The problem of systematic errors has taken on increased importance politically. A Congressional hearing in 2014 about the reproducibility of results specifically addressed systematic errors.

that we develop techniques that can effectively reduce systematic errors.

There are numerous kinds of systematic errors present in galaxy surveys, some known and some yet to be discovered. Extinction and foreground reddening is a common example, but there are others. For instance, SDSS measures of galaxy clustering are biased by an inability to uniformly sample objects due to a physically imposed minimum distance between fibers in the spectrograph (Blanton et al., 2003). Magnitudes can become biased due to edge effects on the Sloan camera that induce blurring and affect image quality especially for point-like objects. Any property calculated through binning can become biased due to improperly handled effects near the edges of the bins.

Other galaxy surveys like the 2-degree-Field Galaxy Redshift Survey (2dFGRS; Colless et al., 2003), the WiggleZ Dark Energy Survey (Drinkwater et al., 2010), the Large Synoptic Survey Telescope (LSST, Ivezic et al., 2008), and Planck (Tauber et al., 2010; Planck Collaboration, 2011) are also susceptible to such effects, making it increasingly important to have a generalized method by which systematic errors can be accounted for and minimized.

In this dissertation, we demonstrate that if fiducial models of the signal and noise covariances are available, it is possible to use Bayesian inference to solve for the most likely value of the signal given the data. We show that if both the signal and noise are Gaussian, then a simple analytic solution for the signal estimate exists in configuration space. We verify this result empirically using a Metropolis-Hastings driven Markov Chain Monte Carlo (MCMC) process. We further argue that if either the signal or noise is non-Gaussian, the

MCMC method can be modified to accommodate other distributions.

We apply this technique to the problem of systematic photometric calibration errors in the SDSS. Because of the rate at which the percentage chance of *actually observing* a galaxy that is *actually present* in a magnitude-limited survey decreases with distance, small offsets in calibration will modulate the number of galaxies observed above one's brightness limit. Nearby volumes possess a large signal component while distant volumes have a substantial Poisson distributed shot noise, plus the modulation of the random photometric effects.

Complicating matters even more is that the galaxies themselves are not uniformly distributed. Instead they possess an intrinsic two-point correlation function which describes the probability of one galaxy existing at a distance r from another galaxy as averaged over the entire sky. This relationship between galaxies is often referred to as "spatial clustering" after the observation that galaxies and groups of galaxies tend to congregate together in particular ways, i.e. along filaments with clusters at the nodes and voids in between, the full picture of which we refer to as the "cosmic web." This scale-dependent clustering is frequently quantified through a statistic known as the power spectrum $P(k)$.

The accuracy of galaxy samples is important, since we now know that they are biased tracers of mass with different types of galaxies possessing different bias factors (e.g. Zehavi et al., 2011; Park et al., 1994). Measuring the distribution of different subsets of galaxies can shed light on mass-galaxy biasing (e.g. McDonald, 2006; Matsubara, 2008; Jeong & Komatsu, 2009).

A critical component of such measurements is the handling of galaxies without spectroscopic redshifts. Because redshifts are the most direct measurable associated with radial distance, any description of galaxy clustering is necessarily incomplete without accounting for these objects. This problem is particularly acute in the SDSS DR6 where approximately 20% of Main Galaxy Sample targets lack spectra.

In Chapters 6 and 8 we test our assumptions about the proper way to count targets without spectra in discretized volumes. We use real Sloan data to simulate mock Universes in which galaxies with known spectra are stripped of that information. Using a combination of discrete counting techniques, probabilistic modeling, and simple scaling, we present optimal counting strategies as a function of volume size, distance, and region characteristics.

While we will frequently invoke the power spectrum as a diagnostic of these data processing methods, we note that the goal of this dissertation is *not* to provide a definitive measurement of $P(k)$ (as Percival et al., 2007, have done, for example) or to offer a best set of cosmological parameter estimates. Rather, we show, in principle, how our framework of signal estimation can enhance derivations of such quantities. As such, this dissertation will not address effects such as foreground extinction (e.g. Schlafly & Finkbeiner, 2011; Schlafly et al., 2010; Schlegel et al., 1998) or redshift-space distortions (e.g. Raccanelli et al., 2013; Szapudi, 2004), which we acknowledge are likely as equally important as the particular photometric effect we seek to minimize.

1.2 Dissertation Structure and Philosophy

In Chapter 2 we provide background information on the SDSS's photometric and spectroscopic pipelines. We review the properties of its geometric regions and query system. We discuss the galaxies on which we perform our analysis and parameterize several functions critical to noise structure and removal.

Chapter 3 provides background information regarding large scale structure including the two-point correlation function and the power spectrum. We connect the two and quantify how they are affected by redshift-space distortions.

In Chapter 4, we explain the need to divide space into tens of thousands of discretized volumes in order to study signal and noise structure. We explain how to simulate mock Universes and generate realizations of clustering signal, shot noise, and systematic zero-point noise. The chapter includes with a summary of notation that is used throughout the dissertation.

In Chapter 5 we delve deeper in the SDSS's photometric and spectroscopic footprints. We describe the manner in which footprint errors can be visualized and discovered. The second half of the chapter focuses on the problem of undersampled regions. The statistical properties of these areas differ so greatly from the rest of the survey that they warrant special attention.

The problem of counting galaxies in cells is covered in Chapter 6. The chapter begins with a census and explanation of nine counting techniques that we test in an effort to account for targets without spectra. Next we explain how and why we split undersam-

pled regions into three types — interspersed regions, dark , and external regions. Through Monte Carlo simulations we demonstrate how each of these region types deserves special handling. We conclude by reporting optimal counting strategies under a variety of conditions.

In Chapter 7 we provide a novel solution for handling statistical and systematic errors in large data sets. We prove that analytic solutions exist for expected signal and noise and provide a Monte Carlo verification and extension algorithm. We demonstrate the merits of this work by estimating clustering signal and power from simulations of contaminated data. We conclude by presenting results from our simulations of shot and systematic noise.

The main text of this dissertation concludes in Chapter 8. Here we combine the results of all previous chapters to clean real data from the SDSS DR6 Main Galaxy Sample. We quantify how the densities of volumes of space change as a result of cleansing. We also provide measures of recovered power and the improved two-point correlation function.

The reader may notice that some sections of the thesis provide more background information than is strictly required. Some of the theory could justifiably have been excluded with the expectation that the content is generally understood by many in the field. Other content could have been referenced in the literature without much additional explication.

When we do include such additional content, the reasoning is deliberate. In some cases it is because we feel that topics covered in various locations ought to be synthesized in a common location and within a coherent, continuous prose. In other cases, particularly as it relates to the Sloan Digital Sky Survey, we have concluded that the background provided

in the literature is insufficient and demands additional explanation. Some geometry issues, for example, were only addressed in private email communications and never published or posted in a clear manner. We suspect that ignorance of the details may explain why errors in the SDSS footprints have gone undiscovered for so long.

When tangential to the main argument, we relegate this information to the appendices. Here one will find topics related to distance measures in cosmology, coordinate transformations, SQL queries used in SDSS data retrieval, derivations of certain power spectra equations, methods for empirical area and volume estimations, correlated shot noise equations, and a method to speed up matrix computations.

Appendix G describes what we refer to as "failed methods." These noise reduction techniques, which largely depend on the limited dimensionality of the zero-point noise, initially displayed much promise. While they ultimately proved inadequate for this particular problem, we believe these methods may be useful in other applications. For this reason (and to avoid letting these substantial efforts go to waste) they are summarized at the end.

Feel free to skip these sections if you are already familiar with the following concepts:

§2.1.1–§2.1.3: SDSS imaging and spectroscopic systems

§2.1.4: photometric redshifts

§2.3.1–§2.3.2: the SDSS Main Galaxy Sample

§2.3.3.1: the Schechter luminosity and selection function parameterizations

§3.1–§3.3: the galaxy power spectrum and two-point correlation function

§3.4: redshift-space distortions

§4.3: principle component analysis

§4.4–§4.5: methods to generate realizations of mock signal and noise

Chapter 2

The Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS; York et al., 2000) is one of astronomy's premiere projects. The SDSS, which began observations in 2000, is a multi-fiber imaging and spectroscopic survey purposed with measuring the positions and properties of hundreds of millions of celestial objects over a large fraction of the sky. A fully digital observatory, the SDSS features a number of innovations that make it the most robust large-scale survey to date.

Through the 1970's, astronomical imaging surveys were often conducted using photographic plates. Light sensitive chemical emulsions would be distributed on a glass substrate, then exposed to the night sky. That technique was both expensive and time consuming.

The advent of digital CCD detectors, which the SDSS employs, offered numerous improvements. Unlike photographic plates, CCDs are linear detectors of light and permit

more accurate measurements of light fluxes over a wider range. Since its data are digital, they can be stored, transmitted, and analyzed far more efficiently and accurately than through past methods.

Sloan's digital technologies, which include both its suite of on-site digital detectors and the computational backbone that supports data analysis and retrieval, has been nothing short of a revolution. Since beginning operations in 2000, the SDSS has photometrically imaged about 500 million objects and taken spectroscopy for about 1 million of those that satisfied certain color and brightness (i.e. apparent flux) criteria. For comparison, the largest single redshift survey prior to Sloan contained on the order of tens of thousands of redshifts while the entire astronomical literature contained about 100,000. This order-of-magnitude improvement has provided so much information about the make-up of our Universe that some have dubbed it the "cosmic genome project."

The SDSS was initially funded for a five-year period, a phase that later became known as SDSS-I (2000–2005). Its primary goal was to image a contiguous, well-calibrated 10,000 $\deg^2$ -sized area of the northern Galactic cap with follow-up spectroscopy of brown dwarves, supernova, Main Galaxy Sample (MGS) galaxies and Luminous Red Galaxies (LRGs). (The latter two categories are discussed in the pages that follow.) The complete catalog of these results is referred to as the Sloan Legacy Survey.

About once a year the SDSS issued a data release (hereafter: DR) containing the latest results with each new release subsuming earlier versions. Through DR5, the SDSS was well on its way to meeting its goal—217 million unique objects over 8000 $\deg^2$ had been

imaged and spectra had been collected for about a million of those.

Sloan's funding was renewed for another three years and the project entered its second phase, SDSS-II (2005–2008). This additional time enabled SDSS to complete the imaging and spectroscopic work begun in SDSS-I. With the release of DR6 in July 2006, the imaging for the Legacy Survey was substantially complete—230 million unique objects had been imaged over an area of 8417 deg^2 (Adelman-McCarthy et al., 2008). Spectra had been taken for 790,220 of those over an area of 6860 deg^2 . A view of the DR6 photometric and spectroscopic footprints are provided in Figure 2.1.

The Sloan Digital Sky Survey's 6^{th} Data Release plays the leading role in the pages that follow for two main reasons. First, while the photometric coverage of the Legacy Survey is largely complete, its spectroscopic coverage contains substantial gaps. The characteristics of these gaps are useful for studying the effect galaxies without spectra have on the number counted within specified volumes.

The second reason DR6 is so valuable is the way that it performs photometric calibrations. It is the final data release that calibrates photometric zero-points in a way that introduces a very distinct sort of systematic error. This process allows me to test a noise reduction technique that cannot be implemented after the improvements of DR7 (Abazajian et al., 2009).

DR6 was very much a survey in transition. Decisions regarding footprint geometry and spectroscopic coverage were made with an eye towards DR7. That is, the choice to collect spectra in some regions but not others was made with the anticipation that the resulting

gaps would be filled in during the next data release. However, this strategy complicates the measurements of certain quantities, like the expected number of galaxies in select regions of space.

In this way, one cannot study DR6 effectively without also understanding DR7. It turns out that some of the geometrical regions that define areas of DR6 are only specified in DR7. I will show a number of cases where this occurs and offer suggestions for improvement.

After the final spectroscopic observations of DR7, the Legacy Survey was effectively complete with a few exceptions. DR8, which marks the beginning of SDSS-III, improves existing data by reprocessing all SDSS-I and SDSS-II imaging data through a new pipeline that better subtracts sky background. It also makes small changes in photometric calibration (an issue I expound upon below) and derives new photometric redshifts. So while DR8 does not introduce any new objects, I will occasionally use its improved measurements to better inform my analysis of DR6.

It is arguable that the legacy of the SDSS may not lie in the astrophysical data it gathered and continues to gather, but in the computational revolution that it inspired. At the time of Sloan's inception, it was whimsically observed that the fastest way to transport a night's worth of data from the Sloan telescope in New Mexico to the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts was by Fed-Ex.

There were a number of challenges that required solutions. The Sloan collaboration, first and foremost, wanted to make the data available to the widest possible user community in an efficient way. This was not only a practical requirement for scientists, but also a strict

requirement of the National Science Foundation (NSF) from which the SDSS derived a portion of its funding. The NSF required that the Sloan data be made public. Figuring out how to do that with the technology available in the early 2000's was indeed a challenge.

The solution to this problem and the technological revolution that it inspired is explored in great length in the book *A Grand Bold Thing* (Finkbeiner, 2010), which I briefly summarize. Ultimately, the SDSS changed the scientific paradigm from "bringing data into a program" to "building a program to bring to the data." The ability to process very large data sets (i.e. "Big Data" in popular parlance) launched a new wave now referred to as *e-Science*.

The collaboration decided to store the survey's raw data in the Data Archive Server (DAS) at Fermilab. From here astronomers were able to access them in the form of FITS images and other binary files. The Catalog Archive Server (CAS) was a companion to DAS and hosted object catalogs, processed data, functions, procedures and other related products in a SQL database. Throughout, I use "CAS" and "the database" interchangeably.

In the beginning, the data was stored in an object-oriented database run by the company Objectivity. Scientists like Ani Thakar, Doug Reynolds and Peter Kunszt worked diligently to create web applications that would optimize searches of the data. When the object-oriented database produced performance problems, downtime, crashes and most damningly, errors in the data, the decision was made in January 2002 to switch to a relational database system run on Microsoft SQL Server.

The relational database system proved more stable and a web portal to the CAS called

SkyServer was created (SkyServer, 2008; Thakar et al., 2008). While CAS existed on the database server, SkyServer existed on a web server. It allowed the user to access data through the Structured Query Language (SQL). Instead of presenting data as objects, it arranged and stored data in about 100 tables.

Once the CAS database grew to 500 GB, SkyServer was no longer able to handle the volume. Queries that returned over 10,000 rows of Sloan data routinely failed. In response, scientists at Johns Hopkins University launched one of its most powerful tools—CasJobs (CasJobs, 2015). CasJobs is a set of XML web services that provide a standardized way to access Sloan data. For the first time it allowed scientists to submit their jobs in an asynchronous way. An SQL query could be submitted to SkyServer where a scheduler would decide which jobs to execute and when.

CasJobs introduced other innovations. It allowed users to have their own data spaces called MyDB that allowed them to share their databases with others. Users were allowed to upload their own data to incorporate with Sloan data. SQL procedures defined in CAS were made openly available. Finding Chart, a tool not unlike Google Maps, was introduced to allow users to visually browse regions of space for galaxies, stars, and other objects.

Of the many tables in the database, a few deserve special mention. The first of these is named `PhotoObjAll`. This table contains all objects' imaging data including unique identifiers, spatial information, fluxes, brightness profiles, when the data were collected, and more. The second table of interest is `SpecObjAll`, which contains all the spectroscopic information gathered during the survey including unique identifiers, links to the

objects' photometric data, redshifts, and spectral classification. The `HalfSpace` table contains the boundaries (i.e. geometry) of SDSS's many regions. Several tables of derived photometric redshifts are available as well.

The database also contains a number of "views". Views don't contain any new information but organize and process them in a way that is more useful to the end user. Examples include `SpecObj`, which filters `SpecObjAll` of duplicate and bad data, and `PhotoPrimary`, which contains only the *primary* imaging data of an object and not data from secondary or repeat observations of that object.

It should be assumed that any data referenced herein were drawn from the SDSS DR6 database and accessed through SkyServer unless otherwise noted. This should allow readers to replicate, verify, and expand upon these results at their inclination.

2.1 Photometry and Spectroscopy

At the start, SDSS operated through the use of two main telescopes located at Apache Point Observatory in Cloudcroft, New Mexico and one offsite calibration telescope. The primary instrument is a dedicated 2.5 meter, 3° wide-field telescope with a modified Ritchey-Chrétien design and exceptionally low distortions along the focal plane (Gunn et al., 2006).

The main telescope can be fitted with both photometric and spectroscopic instruments. The photometric imaging camera (see Figures 2.2 and 2.3) is comprised of six identical, physically separated, horizontally aligned columns of cameras. Each column contains five

two-inch square, 2048 by 2048 pixel CCD detectors, each with essentially nonoverlapping bandpass filters (u', g', r', i', z') (3551, 4686, 6165, 7481, and 8931 angstroms) covering a range from the ultraviolet limit for Earth's atmosphere at 3000 $\mathrm{\AA}$ to the sensitivity limit for silicon CCDs at 11000 $\mathrm{\AA}$ (Fukugita et al., 1996) (see Figure 2.4) for a total of 30 cameras in all (Gunn et al., 1998). To reduce thermal noise, each camera column is sealed in a vacuum and cooled with liquid nitrogen to -80° Celsius. At the time of its introduction, this camera was perhaps the most sophisticated in the word.

The main telescope's spectroscopic instrument is comprised of two fiber-fed double spectrographs in the form of a circular metal tile. Small holes are drilled into the metal tile corresponding to the positions of objects destined for spectroscopic observation. Fibers are subsequently inserted into those holes.

The remaining two telescopes were used through DR7 as calibration devices for the main telescope. These are Apache Point's 20-inch Photometric Telescope (PT), the United States Naval Observatory's (UNSO's) 40-inch telescope in Flagstaff, Arizona (Smith et al., 2002). Each is discussed more in §2.1.2.

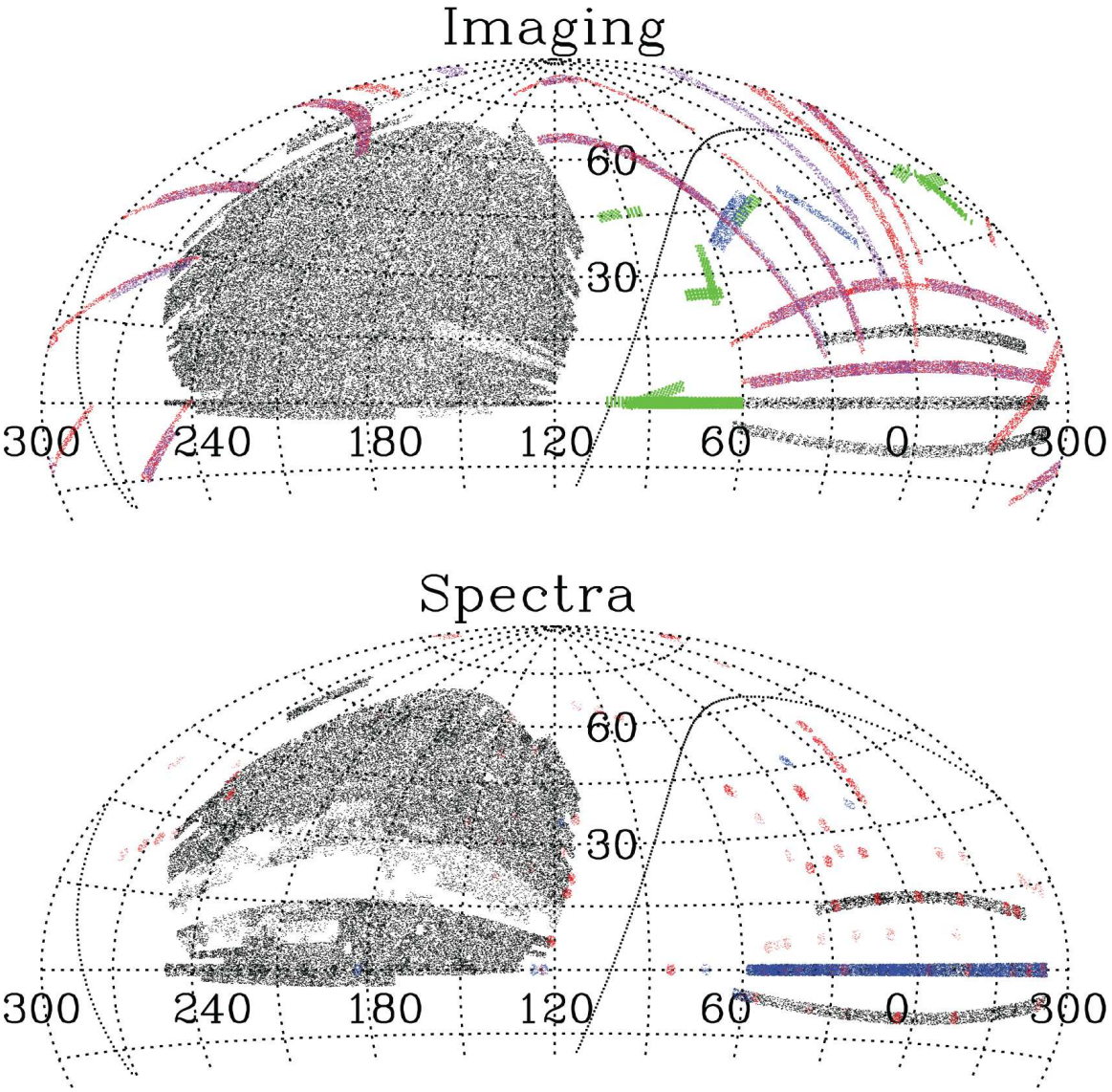

Figure 2.1: A view of the SDSS DR6 imaging and spectroscopic footprints. Areas in gray mark the footprint of the Legacy Survey. Areas in light gray mark the areas new to DR6. Three STRIPEs in the southern Galactic cap are also part of the overall Legacy Survey. The sinuous line between the northern Galactic cap and the STRIPEs in the southern Galactic cap denotes the Galactic plane. Areas in red, blue and green represent other areas observed in for DR6 for projects unrelated to the Legacy Survey. This map is presented in units of right ascension and declination in J2000.0 equatorial coordinates. This figure was first published in Adelman-McCarthy et al., 2008.

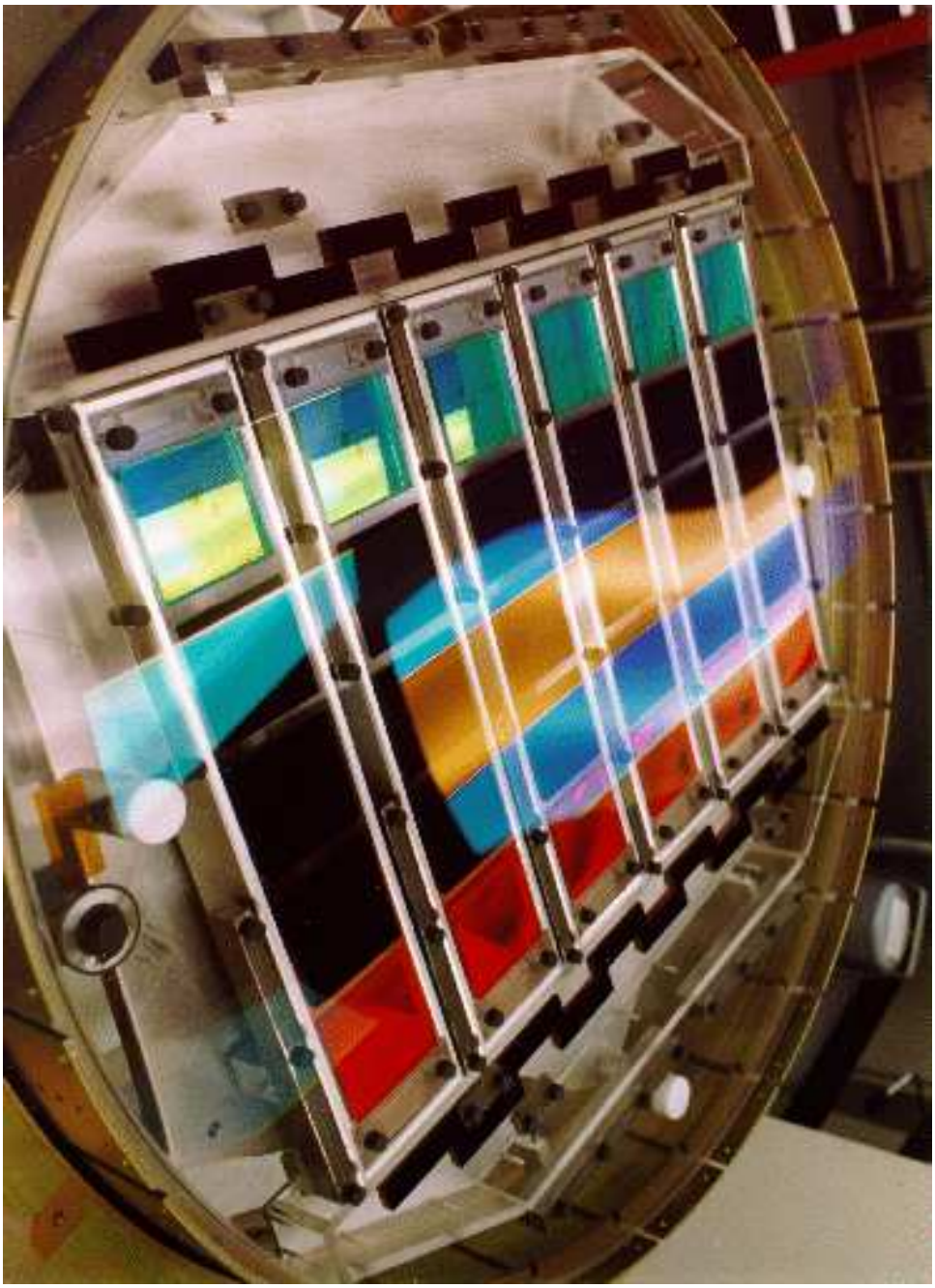

Figure 2.2: This photograph of Sloan's photometric imaging system includes six columns of cameras, with each camera possessing one of five bandpass filters.

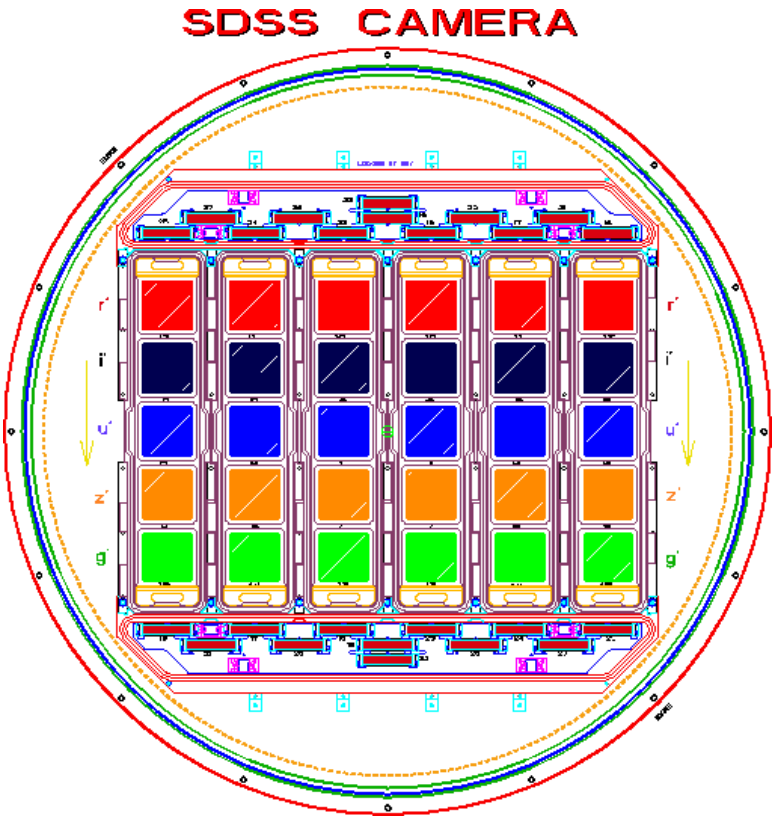

Figure 2.3: Architectural design of the SDSS photometric camera system showing the bandpass filter assigned to each camera.

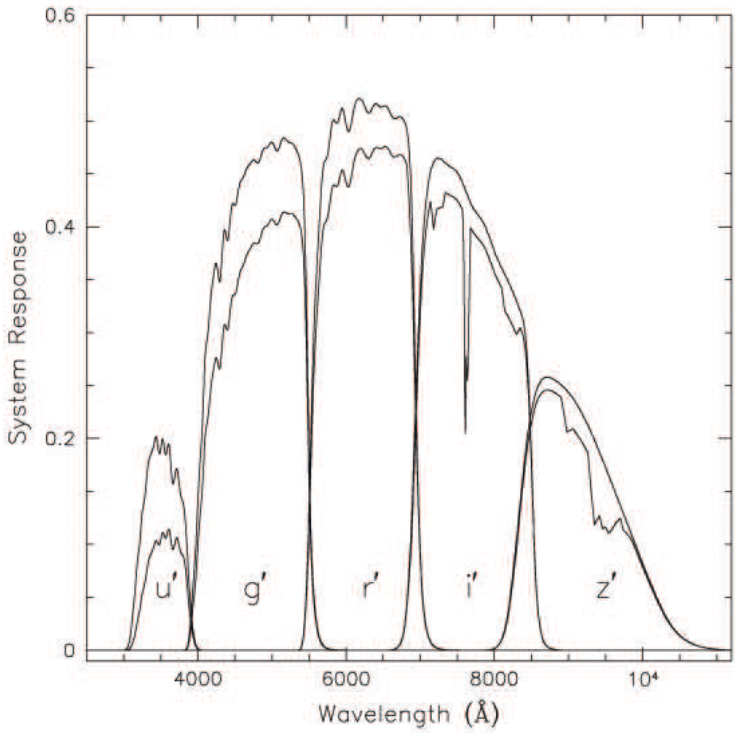

Figure 2.4: System response of Sloan's five bandpass filters. Upper curves show responses without atmospheric extinction. Lower curves show responses with 1.2 airmasses of extinction. Response curves take combined quantum efficiencies of the camera and telescope into account.

2.1.1 Imaging

During the Legacy Survey, the SDSS's main telescope used a *drift scanning* technique. During an observing run, the telescope was pointed and parked. As the sky rotated overhead, light cascaded onto the detectors, producing up to 200 GB of data in a single night.

Each of the six camera columns, or *camcols*, mapped the sky in great circles 10 to 12 arcminutes wide and up to 130° long. These six nonoverlapping *scanlines* collectively formed a *strip*. A second strip, slightly offset from the first, filled in the areas between scanlines and formed a contiguous *stripe* of width 2.5° .

The Legacy Survey primarily observed galaxies in the hemisphere containing the north galactic pole, otherwise known as the *northern galactic cap*, though three additional stripes were surveyed in the southern galactic hemisphere. To gather exceptionally high-quality photometric imagery, observations stayed within 60° of the galactic poles to minimize dust, star fields and atmospheric extinction. Even then, photometric scans occurred only during the best seeing conditions, or about 1.5 days per month on average (Hogg et al., 2001).

Scans from each CCD were split into a series of 2048×1498 pixel *frames*. The set of five frames in a camcol covering the same region of sky is referred to as a *field*. Fields overlap by 128 pixels to ensure objects were not missed. This guaranteed that many objects were imaged multiple times.

Imaging data were processed through software packages known as *pipelines*, two of which deserve particular attention. The first is an astrometric calibration package (Pier et al., 2003) that used two catalogs of astrometric standards to accurately determine the

positions of objects. The precision of these positions depends upon the accuracy of the catalogs and fundamental measuring precision of the telescope.

The second is the *photo* software pipeline which calculated the magnitudes of objects in each band (Lupton et al., 2001). Galaxies have nonuniform light profiles and usually lack sharp edges. For this reason, *photo* returned three separate magnitude models—point spread function (PSF), exponential and de Vaucouleurs—generally used to characterize light from point sources (e.g. stars), disk galaxies and elliptical galaxies, respectively. For objects with measured spectra, *fiber magnitudes*, or the measures of flux within the aperture of a $3''$ diameter spectroscopic fiber, were provided.

The *photo* pipeline contained no prior knowledge of a galaxy's morphology. The optimal magnitude model was selected by comparing likelihoods of the exponential and de Vaucouleurs fits in the r-band. The "winning" model was subsequently applied to the other four color bands. These are referred to as *model magnitudes* and are robust for selecting samples by color.

Measurements resulting from the imaging pipeline's initial run were given the designation *target*. As the SDSS matured, new observations, calibration techniques and improvements to the imaging pipeline were introduced. Objects were reprocessed with the improved measurements being designated *best*.

For the brightest galaxies ($r \lesssim 18$) significant differences could exist between model magnitudes derived between the *target* and *best* phases. These galaxies are better represented by a particularly important product of the *photo* pipelinethe *Petrosian magnitude*.

The Petrosian magnitude measures flux within an aperture that contains a constant fraction of the galaxy's total light. The Petrosian flux (Petrosian, 1976) is defined by

$$F_P = 2\pi \int_0^{kr_P} I(r) r \, dr, \tag{2.1}$$

where $I(r)$ is the surface brightness profile of the galaxy, r_P is a radial distance referred to as the *Petrosian radius* and $k = 2$ is a parameter that defines the Petrosian magnitude to be the total flux within an aperture of radius $2r_P$. The size of the Petrosian radius is chosen such that the average surface brightness at r_P is $\eta = 1/5$ of the mean surface brightness interior to r_P, or

$$\eta = \frac{2\pi \int_{0.8r_P}^{1.25r_P} I(r) r \, dr / \left(\pi\left((1.25r_P)^2 - (0.8r_P)^2\right)\right)}{2\pi \int_0^{r_P} I(r) r \, dr / (\pi r_P^2)}, \tag{2.2}$$

where the values of 0.8 and 1.25 set the size of the annulus used to find the average. The choice of η was deliberate. Too small, and there would be signal-to-noise issues. Too large, and r_P would become sensitive to seeing variations while discrepancies between galaxies fit with exponential profiles versus de Vaucouleurs profiles would become more pronounced. The compromise $\eta = 1/5$ recovered almost all flux for a galaxy with an exponential profile and about 80% of one with a de Vaucouleurs profile.

The Petrosian magnitude is particularly useful in that it does not depend on a galaxy's central surface brightness. It is independent of certain outside effects like cosmological surface brightness dimming, foreground extinction and sky brightness. It is also used to define

the *Petrosian half-light surface brightness*, μ_{50}, which measures (in magnitudes per square arcsecond) the mean surface brightness within an aperture containing half the Petrosian flux.

Finally, if objects were too close together (as detected through light profiles with multiple peaks) the imaging pipeline attempted to *deblend* them. In SDSS parlance, the *blended* parent was deblended into multiple *child* objects.

Each of these photometric characteristics was used determine whether a galaxy would be a member of the Main Galaxy Sample. As the primary galaxy objects of interest within this thesis, their selection criteria will be further explored in §2.3.

2.1.2 Photometric Calibration

The calibration of galaxy magnitudes plays an important role in the analysis to follow. As we will demonstrate in §4.6, systematic errors in the spatially-dependent photometric magnitude zero-points (i.e. deviations from a magnitude error of $\Delta m = 0$) can aggravate the creation of complete galaxy samples which can, in turn, affect measures of galaxy clustering and of all the conclusions that follow. For this reason, we summarize flux measurements and the two distinct photometric calibration methods Sloan used through DR7.

The brightness of an object is measured through its flux density, often given in units of $\mathrm{erg\,s^{-1}\,Hz^{-1}\,cm^{-2}}$. The goal of photometric calibration is the conversion of measured flux density into apparent magnitude in the standardized AB reference system. More than a simple mathematical transformation, this translation requires consideration of the photon-

weighted effective wavelengths of each bandpass filter, the response and ambient environment of the CCD detectors, telescope and airmass transmissions, and instrumental artifacts (e.g. Stubbs & Tonry, 2006).

However, the SDSS cannot observe the true apparent fluxes of objects because intervening atmospheric gases both distort and absorb incoming radiation. The issue is further complicated by the fact that each night the atmosphere is a little bit different. The implication is that for each camcol on each observing run, there exists a small, but non-zero systemic calibration error on the order of a few percent of the flux.

SDSS scientists did their best to address this issue through the use of two photometric calibration methods. The first, which we will refer to as the *PT method*, was used for all data releases through DR7[1]. By utilizing a set of well-understood standard stars, the PT method determined each night's atmospheric extinction and related the measured magnitudes to the stars' uniform photometric system.

This was a multistep process facilitated by 3 telescopes: Apache Point's 20-inch Photometric Telescope (PT), the United States Naval Observatory's (UNSO's) 40-inch telescope in Flagstaff, Arizona (Smith et al., 2002), and the SDSS main telescope. While the intent was to maintain a single filter system throughout, differences in observing environments (USNO: ambient, PT: dry air, SDSS 2.5 meter: vacuum) caused a non-uniform change in the filters' refractive indices. We distinguish these states by referring to magnitudes in the USNO system as primed $u'g'r'i'z'$ and magnitudes in the SDSS 2.5 meter system as

[1]PT calibration is the default calibration for all releases DR1 through DR6 and is directly incorporated into the reported magnitudes. PT calibration information for DR7 is stored separately in CAS tables `OrigField` and `OrigPhotoObjAll`.

unprimed $ugriz$.

Absolute calibration required comparisons with well-studied standards, typically bright stars. Because these objects tend to be brighter than the SDSS main telescope's $r = 14$ saturation limit, their observations were conducted by the less sensitive UNSO telescope.

Over the course of two years, UNSO repeatedly observed a set of 158 bright primary stars (Smith et al., 2002) drawn from the Northern sky. The stars were selected to span a range of colors, right ascensions, airmasses (to quantify atmospheric extinction) and brightness between $r = 8$ and $r = 13$. They were linked to an absolute flux system via a single F0 subdwarf star BD+17-4708 whose flux in the SDSS filters is well understood (Fukugita et al., 1996).

A set of 1520 $41.5 \times 41.5\,\mathrm{arcmin}^2$ transfer fields, referred to as *secondary patches*, were geometrically positioned throughout the survey area such that a set of four spans the width of a full stripe. During an observing run, the PT observed the primary stars along with the secondary patches that overlapped that night's SDSS main telescope run. The secondary patches were first calibrated to the photometric standards in $u'g'r'i'z'$ then transformed to $ugriz$. These operations were handled by two data reduction pipelines the Monitor Telescope Pipeline (mtpipe) and the Final Calibrations Pipeline (fncalib) (Tucker et al., 2006).

Ultimately, the errors which accumulated during this process were primarily the result of the slightly different photometric systems and unmodelled atmospheric variations at Apache Point Observatory. These contribute to an overall photometric uncertainly of

approximately 2% rms (Stoughton et al., 2002; Abazajian et al., 2003, 2004) and 2% in r (Ivezić et al., 2004).

The second calibration method, referred to as "ubercalibration" or *ubercal* (Padmanabhan et al., 2008), used repeat observations of previously calibrated SDSS data to increase consistency. Such data came from overlapping camcols and the "Apache Wheel scans," which were run perpendicular to the main survey stripes. This method yields residual errors of about 1% in $griz$ and 2% in u. Ubercal is the de facto calibration of DR7[2].

In §4.6 we use the ubercalibrations from DR7 to estimate that the distribution of PT calibration photometric zero-points in DR6 is roughly Gaussian with a standard deviation of $\sigma_m = 0.01$. That distribution will parameterize photometric zero-point models that can ultimately be leveraged to reduce the systematic noise present in galaxy density measurements.

The SDSS calibration process possesses both similarities and differences with other major surveys. As with the PT, the Javalambre-Physics of the Accelerated Universe Astrophysical Survey (J-PAS) uses an auxiliary telescope to classify millions of stars to serve as standards (Benítez et al., 2015) for calibrating each exposure. Like ubercal, J-PAS performs follow-up calibrations once 4 exposures are gathered. Mathematically, this is accomplished by constructing a spectroscopic model for the SDSS stellar locus, finding its expected value, and simultaneously solving for the zero-points by fitting to instrumental stellar magnitudes (Kelly et al., 2014).

[2]This so-called *ubercal* approach was first introduced as a standalone table, `UberCal`, in DR6 though it was not incorporated into the reported magnitudes. DR7 does have a separate `UberCal` table, but this is a carryover from DR6 that the database managers forgot to delete and should be ignored.

The SuperNova Legacy Survey (SNLS) compares point spread functions (PSF) and aperture fluxes from non-variable stars that are imaged multiple times (Astier et al., 2013). Using a least-squares solution, the SNLS arbitrarily fixes one zero-point and calibrate the others with respect to it (Desai et al., 2012). The ALHAMBRA survey solves stellar transformation equations (Aparicio Villegas et al., 2010) by referencing common objects in 2MASS (Cutri et al., 2003), SDSS, and the Next Generation Spectral Library[3] and selecting those with higher signal-to-noise (Cristóbal-Hornillos et al., 2009).

LSST, the spiritual successor to Sloan, uses detectors whose relative sensitivities are known to better than 1 part in 10^3. These are then used to measure on-site relative throughputs and normalized using a highly precise calibrated detector. PanSTARRS (Schlafly et al., 2012) and DES calibrate similarly. LSST's survey strategy involves imaging areas of the sky multiple times over 10 years, which also permits a sort of self-ubercalibration (Ivezic et al., 2008).

2.1.3 Spectroscopy

Mapping the three-dimensional distribution of galaxies is of primary importance to those seeking to understand the evolution of large scale structure in the Universe. Such an exercise cannot be undertaken without knowledge of galaxies' radial depths and no method can determine those depths more accurately than gathering spectroscopic redshifts. The suite of spectroscopic tools within the SDSS main telescope was designed specifically for

[3]https://archive.stsci.edu/prepds/stisngsl/

this purpose.

Before spectroscopy can occur, a target selection algorithm must determine whether an imaged object should be a candidate for spectroscopic observation. The Legacy Survey's objects of greatest interest are brown dwarfs, hot standards, quasars (QSO's), LRGs, and MGS galaxies, each of which has its own photometric selection criteria. Because this thesis concerns MGS galaxies, discussion of selection criteria will be limited to this category. Details are covered in §2.3.1[4].

The SDSS main telescope used two multi-object fiber-fed spectrographs (Uomoto et al., 1999). Each has a blue-band (3800-6150 Å) and red-band (5800-9200 Å) camera covering 4098 pixels. The cameras offer a spectral resolution $\lambda/\Delta\lambda$ between 1850 and 2200, depending on the wavelength.

Fibers aligned along two slitheads transmitted light from the focal plane of the telescope to the spectrographs. Each spectrograph could accept a maximum of 320 $3''$ diameter fibers for a total of 640.

At the focal plane, the fibers were manually inserted into holes in a circular disk known as a *tile* (see Figure 2.5). Each tile was a 1-meter diameter, $^1/_4$-inch thick circular disk of aluminum which, once inserted into the main telescope, subtended an angular radius of $1.49°$. The position of each hole/fiber on the tile corresponded to the position of a spectroscopic target. Because each region of the sky offered its own unique set of targets, a couple thousand tiles needed to manufactured. The holes were drilled off-site and the tiles

[4]The results of the target selection algorithm are stored as bit masks in the fields *primTarget* and *secTarget* in CAS table `PhotoObjAll`.

were then transported to Apache Point where the fibers were subsequently inserted.

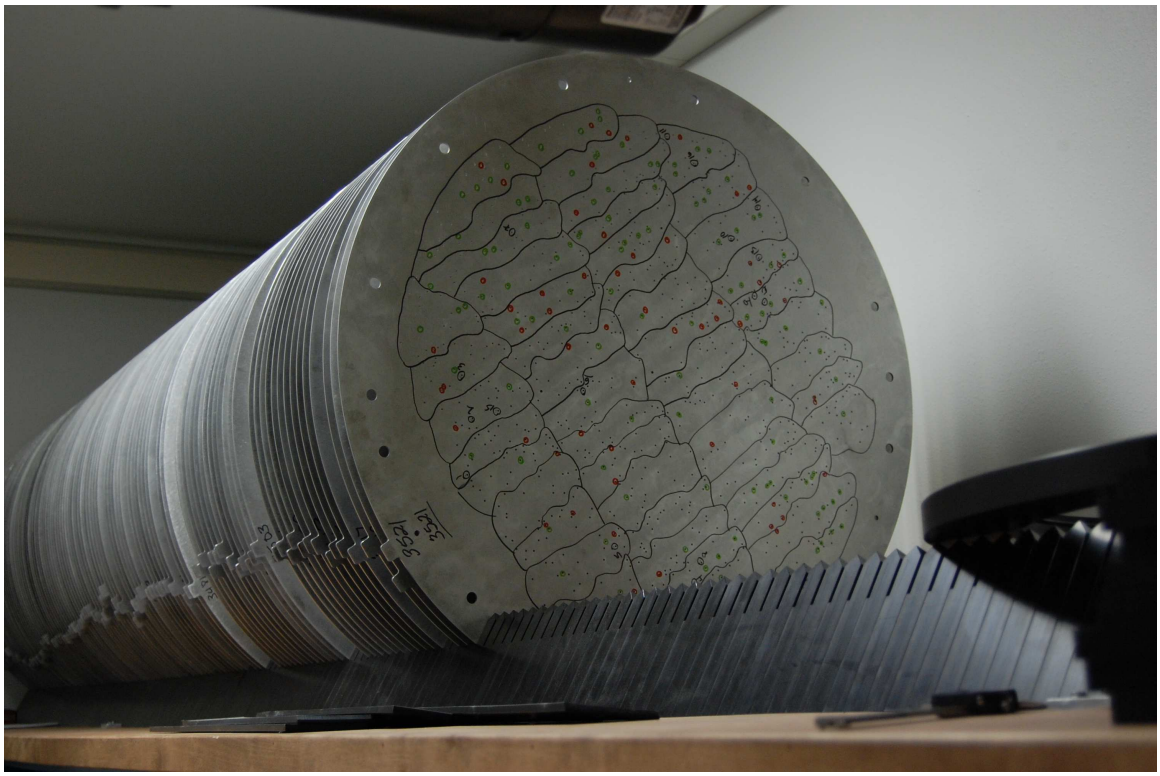

Figure 2.5: Collection of tiles stored on-site at Apache Point.

Of the 640 available fibers, 48 were reserved for observing sky backgrounds and spectrophotometric standards, leaving 592 fibers for spectroscopic targets. Due to the size of their fiber claddings, no two fibers on a single tile could be positioned closer than $55''$ of one another. This constraint is frequently referred to as *fiber collisions*.

The effect of fiber collisions was partially mitigated by the fact that tiles' footprints were permitted to overlap. This permitted pairs of objects with separations less than $55''$ to be spectroscopically observed as long as their respective fibers were placed on separate tiles.

Obtaining spectra is both financially expensive and time consuming. Not every objected

targeted for spectroscopic observation could ultimately have its spectra taken. Deciding which targets would be assigned fibers was accomplished through the use of the *tiling algorithm* of Blanton et al. (2001).

This algorithm identified all spectroscopic targets within a rectangular area called a *chunk* and prioritized them based on type. Brown dwarfs, by virtue of their rarity, received highest priority followed by QSO's, then MGS galaxies and LRGs. The latter two types received equal treatment, meaning that when fiber placement between the two came into conflict, the target was selected essentially at random.

Once targets were selected, tiles were projected nearly uniformly on the sky. To determine which targets to assign to each tile, a maximal set of targets separated by at least $55''$ was identified. This became known as the *decollided set*.

Next, the algorithm iteratively perturbed the centers of the tiles with the goal of maximizing the number of quality spectra. A cost function was applied and minimized until an optimal orientation was found that assigned fibers to $> 99\%$ of decollided targets.

After each of these *tiling runs* was performed, spectroscopic observations could begin. These occurred in 15 minute bursts and continued until the signal-to-noise was $S/N > 4$ for objects with fiber magnitudes > 20.2 in g and > 19.9 in i. This typically took about 45 minutes under good conditions (e.g. dark sky, good seeing), though observations could extend between nights.

Once gathered, the spectroscopic data were reduced through the (publicly available) `idlspec2d` software package (Stoughton et al., 2002). Data from the blue and red cam-

eras were combined and the spectra were analyzed. The two primary data products were the classification of the target (e.g. star, galaxy, etc.) and its redshift.

The data reduction pipeline included models for galaxies, stars, QSO's, and cataclysmic variables in the form of their eigenspectra (as found through Principle Component Analysis—more on this principle in §4.3) based upon a set of narrow, common nebular emission lines (SubbaRao et al., 2002). Each target's spectrum was compared against these model eigenbases at a given redshift and evaluated through a chi-squared goodness of fit. The eigenbases were then incremented to a larger redshift and the process continued.

In the end, the target was given a classification and redshift corresponding to the best overall chi-squared fitting. In some cases, the chi-squared fitting between the best and next-best classification/redshift was too small to make an accurate determination. In such cases, a flag was set in the database to indicate a poorly determined or otherwise inconsistent result. The parameter *zConf* in the `SpecObj` table quantifies the redshift confidence. Repeat observations of spectroscopic targets have found that for galaxies near the flux limit, accuracy of better than 30 km/s has been achieved.

Finally, while it might seem optimal to have tiled the entire sky at once, there are compelling reasons to have done so in *chunks*. Unlike imaging, perfect seeing was not required for spectroscopy. While imaging only consumed about 1.5 days/month, the intervening time could be put to good use gathering spectra. Furthermore, it was deemed preferable to release full galaxy information (i.e. including spectra) to the scientific community as quickly as possible.

2.1.4 Photometric Redshifts

Measuring redshifts of dim and distant objects with the Sloan spectrographs involves spreading limited light among approximately 4000 spectral elements. Such splitting leaves limited signal, and that which does remain must compete with sky and instrument noise. This helps explain why spectroscopic runs took the better part of an hour and why the Legacy Survey was only able to gather spectra for less than 1% of detected objects. Despite being prioritized by the tiling algorithm, only 80% of MGS targets had their spectra collected by the end of DR6.

Scientists account for this deficit by estimating galaxy distances using other correlated variables. Perhaps the most popular alternatives are *photometric redshifts* or *photo-z's*. As the name implies, photo-z's estimate redshifts by exploiting correlations between distance and imaging data along multiple channels. They are almost always less accurate and less precise than spectroscopic redshifts, but often constitute the best available option.

Efforts to derive redshifts photometrically date back 50 years (e.g. Baum, 1962; Weymann et al., 1999) and the methods developed can be classified into two broad categories: template-fitting and training-set. In the analysis to follow in future chapters, both types of photometric redshifts will be utilized to "fill in the gap" left by the 20% of spectroscopically unobserved MGS targets.

The template-fitting approach attempted to match an object's multi-color spectral energy distribution (SED) with model or empirical templates of known objects (see e.g. Budavári et al., 2000; Csabai et al., 2003; Coe et al., 2006; Brammer et al., 2008). With

this method, each object was compared against a small library of templates (e.g. elliptical galaxies, Sbc, Scd, irregulars) at various redshifts. The template that maximized the maximum likelihood function was used to assign an object type and redshift. Error estimates were based upon the shape of the χ^2 minimum, but are generally unreliable. Furthermore, reported errors do not take template errors into consideration.[5]

Training-set methods, on the other hand, use empirical spectroscopic and photometric information from known galaxies to train estimators to best predict unclassified objects' types and redshifts. There exist a variety of these methods including Artificial Neural Networks (ANN) (e.g. D'Abrusco et al., 2007; Collister & Lahav, 2004; Vanzella et al., 2004; Tagliaferri et al., 2003, the Nearest Neighbor Method (Csabai et al., 2003), polynomial fitting (Wang et al., 1998; Connolly et al., 1995), the Nearest Neighbor Polynomial (NNP) technique (Lima et al., 2008), Random Forests (Carliles et al., 2010), and Support Vector Machines (Wadadekar, 2005). For a comparison of methods see e.g., Dahlen et al. (2013); Hildebrandt et al. (2010).

For DR6 the training-set approaches of Oyaizu et al. (2008) were incorporated directly into the database table `Photoz2`. Their team used an ANN method trained with 639,911 spectroscopically observed galaxies, all of which possess SDSS photometry. About 83% of the spectroscopic training set came from the SDSS itself, while the remainder was filled out with objects from the CNOC2, CFRS, DEEP, DEEP2, TKRS, and 2dF-SDSS LGR and QSO surveys. The training set endeavored to be robust and representative by spanning the

[5]The template-fitting method was first employed for the SDSS in DR5. In DR6 results are stored in table `Photoz`.

same luminosity, color, and redshift ranges as the SDSS photometric sample.

Artificial neural networks require choosing a particular network structure, then estimating parameter values through gradient descent. There are two sets of photo-z's (and their errors) reported in `Photoz2` CC2 and D1.[6] The CC2 trained using three concentration parameters, which are measures of how tightly clustered a galaxy's light is, and four colors: $u-g$, $g-r$, $r-i$ and $i-z$. The D1 set is trained using five magnitudes ($ugriz$) and five concentration parameters and have been shown to display better performance at brighter magnitudes.

Unlike training-set methods, spectral template methods do not require large training sets. However, the advent of large-scale surveys like SDSS and the forthcoming LSST render this less of an advantage. In fact, comparisons between the methods have shown that with surveys the scale of SDSS, photo-z's derived through training sets exhibit less bias and scatter relative to their true spectroscopic redshifts than those from template methods (Cunha et al., 2009).

In either method, photometric redshifts are conditioned upon the colors of their galaxies and are thus Bayesian reconstructions of their true redshifts. Therefore each photo-z should be considered less of a determined number and more (when combined with its error) of a probability distribution. This statistical interpretation of photometric redshifts will be put into practice in Chapter 6.

Finally, we acknowledge that there are other correlations that could be exploited in

[6]The errors σ_z reported in `Photoz2` represent the 1-sigma or 68% confidence limit. Oyaizu et al. (2008) find that 68% of galaxies in their validation set possess $\sigma_z \leq 0.021$ for D1.

pursuit of better photo-z's (e.g. the effect of galaxy inclinations as reported in Yip et al., 2011). As the purpose of this dissertation is not to present a perfect reflection of the cutting edge of photo-z analysis, these additional effects will not be considered here.

2.2 Geometry

This section reviews the geometric properties of the SDSS photometric and spectroscopic footprints. The footprints, loosely defined, are the areas within which observations have already occurred. In much the same way as a map of the United States can be described by the boundaries formed by state lines, congressional districts, electric interconnections, lakes, sports conferences and many more, the SDSS footprints are defined in terms of "regions", of which there are over 20 types.

The first part of this section focuses on the extent of the imaging survey, or the *photometric footprint*. While basic geometric descriptions might only consider STRIPEs, STRIPs, and CAMCOLs, there are many second-order regions of considerable importance. Some define the difference between what the Legacy Survey intended to observe versus what it actually observed. Others, like PRIMARY SEGMENTs, are not explicitly defined in the database but play prominent roles in the propagation of photometric zero-points.

Next, we review the main principles that define the spectroscopic footprint. These include TILEs, which are the spatial manifestations of the tiling algorithm, and SECTORs, which are formed by the thousands of Venn diagram-like intersections of TILEs and other

masks. We conclude with an explanation of *halfspaces*, which both define regions and facilitate rapid searches within them. In all cases, original visualizations of Sloan geometry are provided to complement the text.

The reader may find this section to be overly detailed, but with good reason. It is my opinion that no comprehensive and coherent explanation of the Sloan geometry exists. This stands as my attempt to create one. Furthermore, as we will argue in Chapter 5, there are myriad errors in the DR6 footprints that adversely affect measurements of large scale structure. This section forms the basis upon which those errors will be corrected.

Some region names, like TILE and SECTOR, have general definitions in addition to their SDSS-defined ones. When referring to SDSS regions in particular we capitalize their names for clarity. Extracting the geometric properties of regions is done through queries of the database. Explicit queries for each region discussed in this section are provided in Appendix C.

2.2.1 Photometric

The total region of the sky imaged by Sloan is the FOOTPRINT, or to distinguish it from spectroscopic coverage, the *photometric footprint*. By virtue of the SDSS's drift scanning approach, the photometric footprint is the union of areas enclosed within sets of great circles. These circles are defined with respect to an imaginary axis that passes through two stationary poles at $(\mathrm{ra}, \mathrm{dec}) = (95^\circ, 0^\circ)$ and $(275^\circ, 0^\circ)$.

During a single drift scan, each of Sloan's six cameras observes its own, nonoverlap-

ping scanline covering an abstract region known as a CAMCOL. The union of those six CAMCOLs is a region known as a STRIP. The region covered by two (slightly overlapping) interlocking STRIPs is known as a STRIPE.

STRIPEs are separated by 2.5° and span from pole to pole. (In practice observations never spanned this full distance.) Each is defined and indexed by its inclination relative to the equator such that a STRIPE of index n has an inclination of $-25 + 2.5n$. For example, STRIPE 10 lies along the equator and STRIPE 11 lies 2.5° above it in the northern hemisphere. The highest possible latitude STRIPE is at $n = 46$, or 90°, though only STRIPEs 1 through 45 are formally defined in the database. In DR6 there are also three STRIPEs—76, 82, and 86—defined in the southern hemisphere.

STRIPEs have their own coordinate system known as "great circle coordinates". If the center line of each STRIPE (1.25° from each boundary) acts as its own equator, then the SDSS coordinates mu/nu act as ra/dec for this truncated region of the sky. While STRIPEs are, in principle, abstract regions spanning pole to pole, they are not defined as such within the database. Rather, they are assigned mu limits that more closely align the abstract ideal of a STRIPE to what was actually observed. The definitions of STRIPs and CAMCOLs are similarly limited.

Closely linked to great circle coordinates are "survey coordinates" as shown in Figure 2.6. The dimensions "eta" (survey latitude) and "lambda" (survey longitude) are rotations of the main coordinate system where $(\mathrm{ra}, \mathrm{dec}) = (275^\circ, 0^\circ)$ and $(0^\circ, 90^\circ)$ correspond to $(\mathrm{eta}, \mathrm{lambda}) = (0^\circ, 90^\circ)$ and $(57.5^\circ, 0^\circ)$ respectively. More intuitively, lines of constant

"lambda" are great circles centered on the survey poles while lines of constant "eta" are like those of "mu", but are allowed to extend beyond an individual STRIPE.

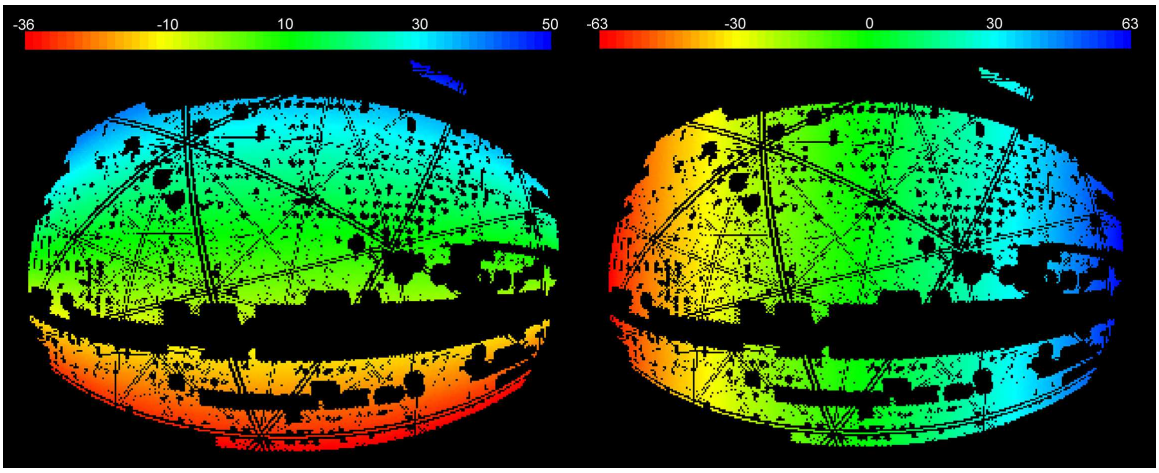

Figure 2.6: SDSS survey coordinates within the DR6 spectroscopic footprint. "Eta" coordinates (*left*) are oriented perpendicular to STRIPEs while "lambda" coordinates (*right*) run parallel to their lengths. The triangular, crisscross patterns correspond to the positioning of spherical cells (see §4.1). They are not reflective of the survey geometry.

There were instances in which the *actual* survey geometry differed from the *idealized* survey geometry. For example, in the early data releases there were slight deviations of a few arcseconds in latitudinal pointing from what was planned. In other cases, the Sloan telescope concluded observing before or after reaching the limit of a STRIPE.

The SDSS database reports the true survey geometry through regions called CHUNKs. While there were only 48 STRIPEs intended for DR6, there are 111 CHUNKS, meaning the average STRIPE was "broken up" into two to three separate pieces during observing runs. If the survey had been conducted "perfectly", then CHUNKs and STRIPEs would have been identical.

Because STRIPEs are rectangular objects projected onto a sphere, they begin to overlap

as they approach the poles. Similarly, CHUNKs overlap as illustrated in Figure 2.7.

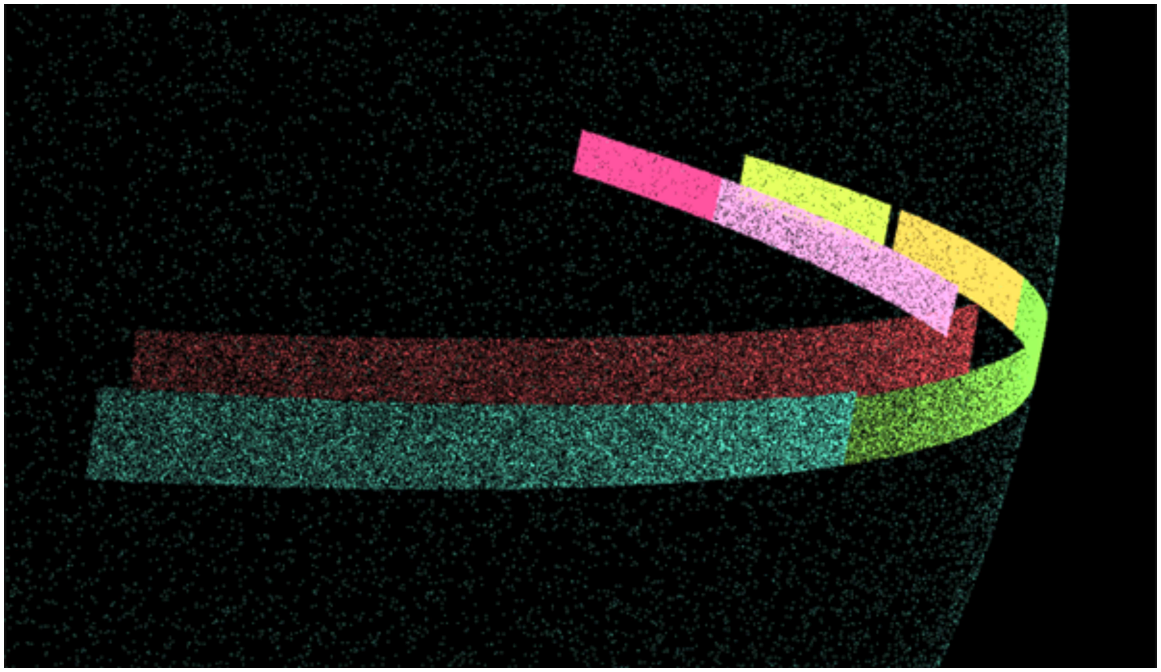

Figure 2.7: A set of seven DR6 CHUNKs from three adjacent STRIPEs are projected onto the celestial sphere. All CHUNKs possess the same angular height, so it is clear from the middle of this figure that the red CHUNK overlaps the teal and green CHUNKs. Assigned colors are random.

Targets observed within CHUNK overlap regions are usually imaged at least twice, once for each scan of the region. Once CHUNKs are resolved by the pipeline, several factors determine which observations will be considered *primary* and which will be relegated to *secondary* or *tertiary* status.

The region that exclusively contains a CHUNK's primary objects is referred to as a PRIMARY. In DR6 there are 111 CHUNKs and therefore 111 PRIMARYs. Each pair shares a unique *chunkID* that can be found in the `Segment` table. No CHUNK's area beyond the limit of its corresponding STRIPE is permitted to lie within a PRIMARY region.

A visualization of CHUNKs and PRIMARYs that extends the example of Figure 2.7 is provided in Figure 2.8. Figure 2.9 offers a full sky view of all DR6 PRIMARYs.

Because this investigation only considers objects with primary status (i.e. those within PRIMARYs), nothing outside the union of PRIMARYs is investigated. For simplicity, hereafter the phrase *photometric footprint* shall be synonymous with *the union of all PRIMARYs*.

Just as each STRIPE is comprised of 12 CAMCOLs, each CHUNK is comprised of 12 SEGMENTs. In this way, SEGMENTs can be thought of as the realized observations of the abstract CAMCOLs. There are 48 STRIPEs which means that under ideal observing conditions, only $48 \times 12 = 576$ distinct CAMCOLs would exist.

Of course, ideal observing conditions are the exception rather than the rule. Due to effects such as the deterioration of seeing conditions during the night, full STRIPEs are rarely observed in a single run. The complete imaging of DR6 required 171 runs, which created $171 \times 12 = 2052$ SEGMENTs. (See Figure 2.10.) As with CHUNKs, CAMCOLs are redefined *in the database* such that their angular limits match those of their corresponding runs.

Just as PRIMARYs are the non-overlapping portions of CHUNKs that contain primary observations, PRIMARY SEGMENTs are the non-overlapping portions of SEGMENTs that contain the same. Figure 2.11 offers a visual description of how SEGMENTs are cropped to form PRIMARY SEGMENTs. Unlike the other regions previously described, PRIMARY SEGMENTs are not explicitly defined in the database. Instead, their geometric

properties are derived by applying the PRIMARY constraints atop the SEGMENT definitions. The query to do so is included in Appendix C. The union of PRIMARY SEGMENTs occupies a combined area of 8304 $\deg^2$, or 98.7% of the 8417 $\deg^2$ covered by the full photometric footprint (Adelman-McCarthy et al., 2008).

Perhaps the most important point regarding DR6's 2052 PRIMARY SEGMENTs is that each one is independently photometrically calibrated. This means there is a fixed but unknown set of 2052 photometric zero-point offsets. Each affects the measured magnitudes of the targets within their boundaries in a unique way. Accounting for the cross-SEGMENT discrepancies this introduces will be a major focus in the pages to come.

2.2.2 Spectroscopic

The area of the sky observed by a physical metal tile is referred to as a TILE region. While tiles can only be inserted into the spectrograph one at a time, TILEs may overlap to increase the effective density of available fibers. The number of TILEs overlapping an area of the sky is referred to as that area's *depth*. Greater depth generally implies greater spectroscopic coverage.

Multiple TILEs are generated during each tiling run. Such runs are contained within *tiling boundaries*. These boundaries are referred to as TIGEOM regions within the database. Parts of the sky for which no spectroscopic observations are desired are covered with *tiling masks*. The area within the tiling boundaries but outside the tiling masks is referred to as the *tiling region*.

DR6 has 1520 TILEs with *regionID*'s between 1839 and 3358, as shown in Figure 2.12. The gap in the equatorial declinations of the northern galactic cap represents an area that was imaged as of DR6, but not yet tiled. Those regions were "filled in" with other TILEs during DR7.

SECTORs are non-overlapping intersections of tile regions. Under the simplest circumstance a SECTOR would be a single circle corresponding to its TILE. In practice, the application of tiling boundaries, tiling masks, and intersections with other TILEs creates thousands of additional intersections, each one of which is its own SECTOR. A visual example of the SECTORs within a randomly selected TILE is provided in Figure 2.13.

There are 9464 distinct SECTORs defined within the DR6 database. Each is provided its own *regionID*. In DR7, new SECTORs were introduced as spectroscopic observations continued. This enlargement of the spectroscopic footprint did not change the definitions of any of the DR6 SECTORs, but it did change their unique indices.

Because spectroscopic observations only occur within SECTORs, hereafter *spectroscopic footprint* should be considered equivalent to the *union of all SECTORs*. As verified by Figure 2.14, the spectroscopic footprint lies entirely within the boundaries of the photometric footprint. There are also a considerable number of "holes" in the spectroscopic footprint as compared to the TILE footprint. This is due to a number of effects including the introduction of tiling masks and differences between the intended and realized spectroscopies.

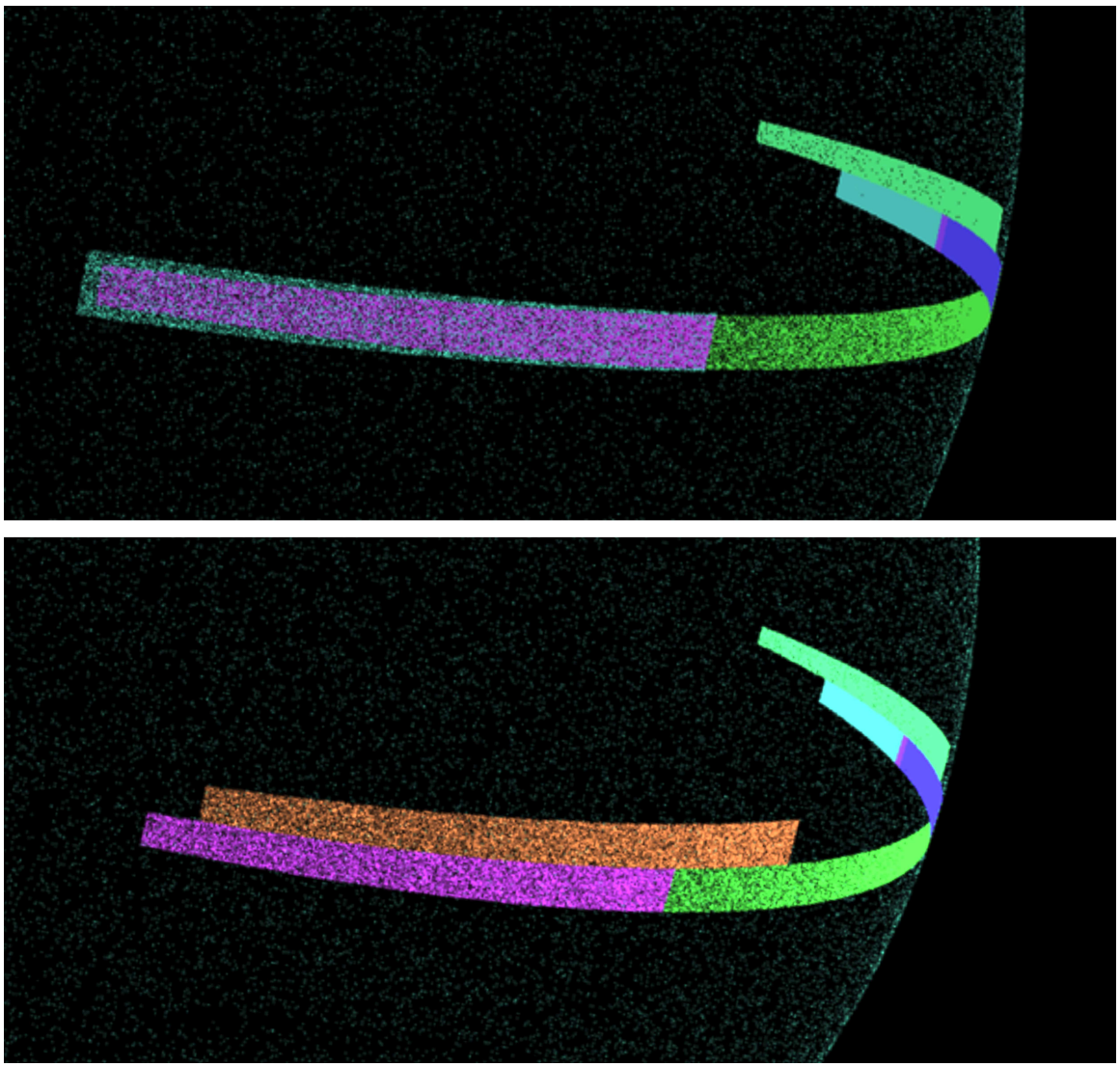

Figure 2.8: Visual representation of CHUNKs and PRIMARYs in the same region of sky. *Top*: A CHUNK (cyan) encloses its PRIMARY (purple). A PRIMARY's area is always less than or equal to the area of its CHUNK. *Bottom*: The non-overlapping PRIMARYs are visualized in random colors.

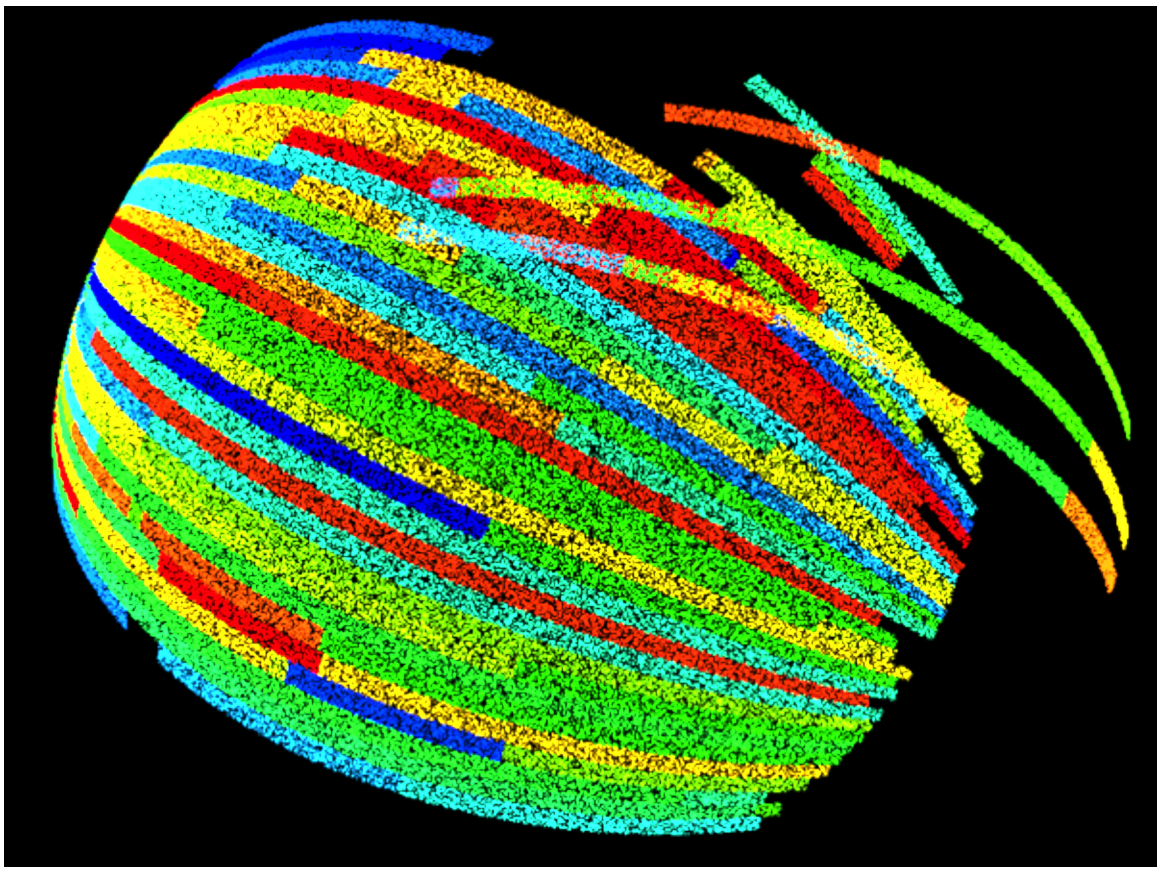

Figure 2.9: Visualization of DR6's 111 PRIMARY regions. No PRIMARYs overlap. Each is assigned a random color to distinguish it from its neighbors. An average of two to three PRIMARYs compose each STRIPE.

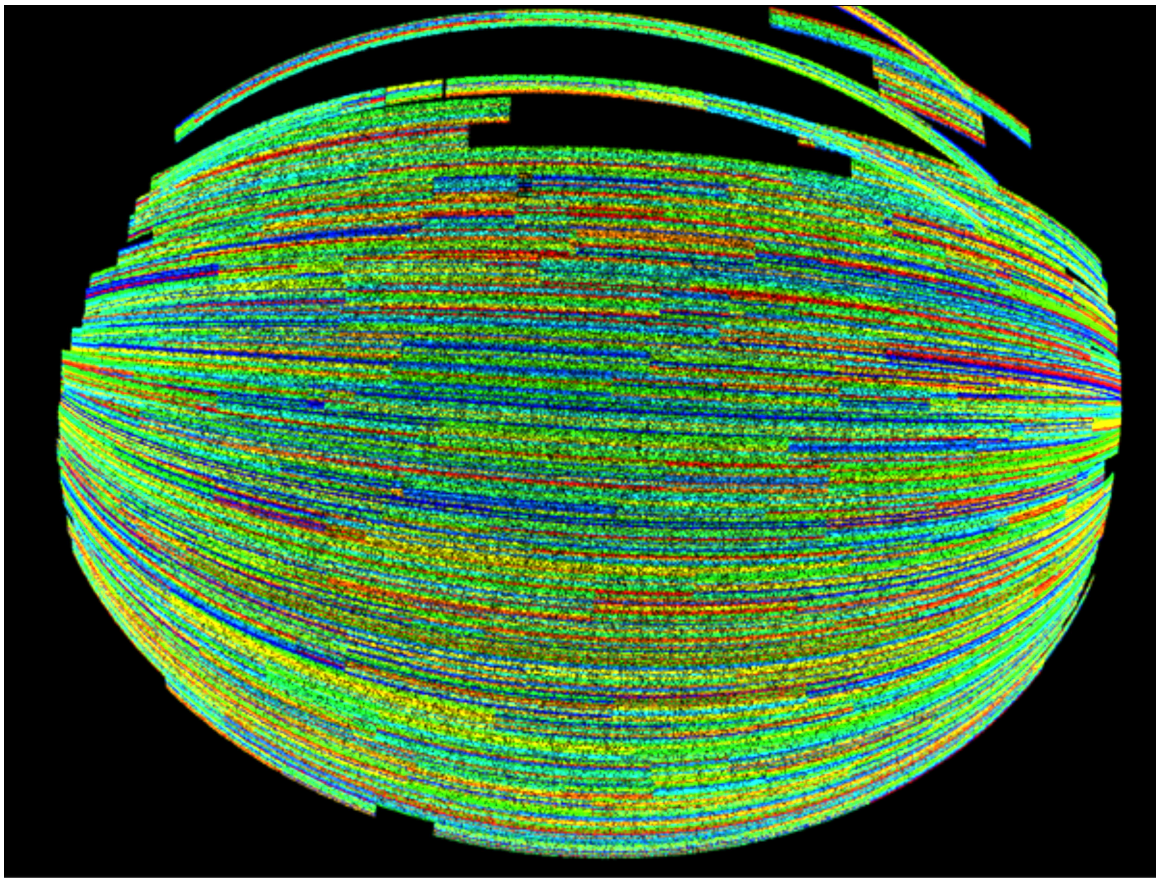

Figure 2.10: Visualization of the 2052 DR6 SEGMENTs. Regions in the northern galactic cap comprise the majority of the image while portions of the three STRIPEs in the southern hemisphere are visible at the top. Each SEGMENT is assigned a random color to distinguish it from its neighbors. SEGMENTs are grouped in sets of 12 such that the angular extent in mu is the same for all. As the SEGMENTs approach the poles, they overlap to a greater degree.

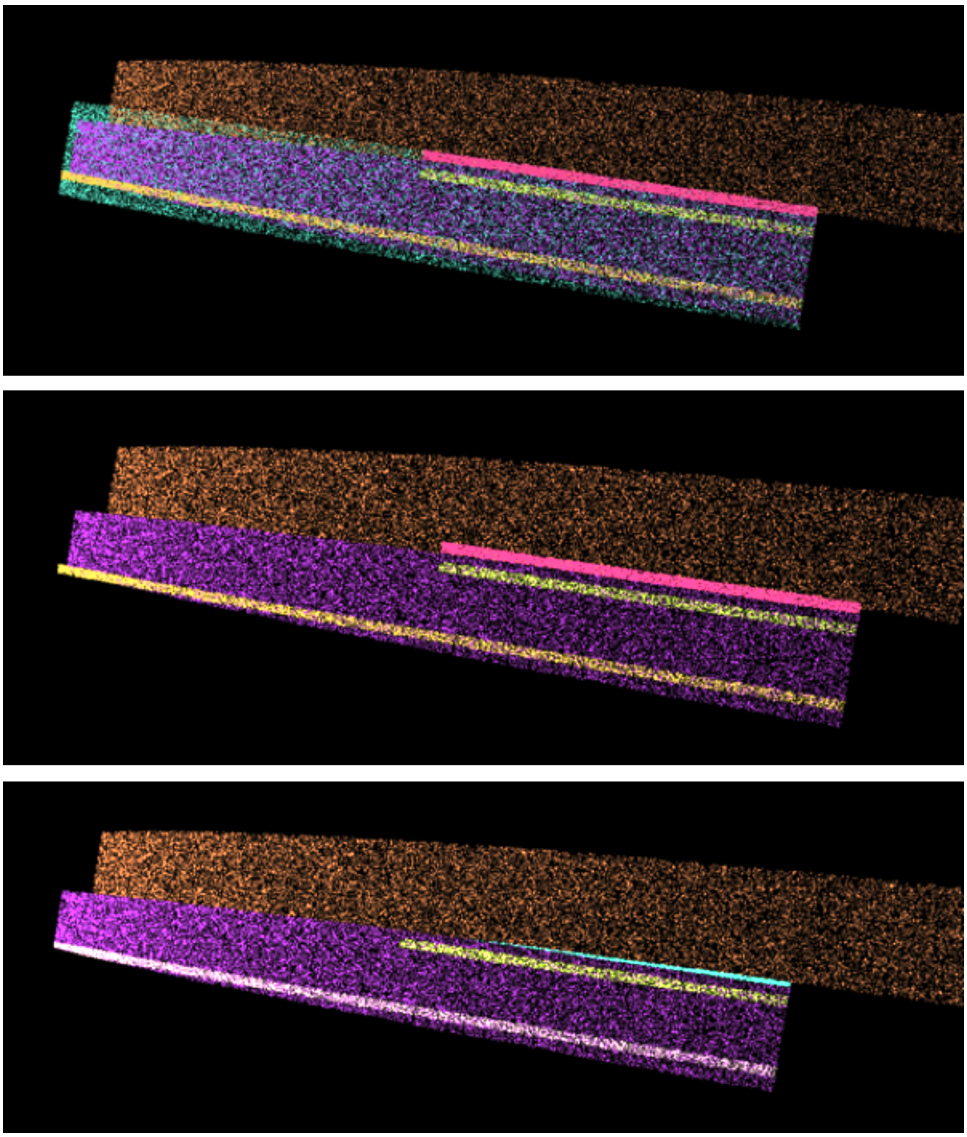

Figure 2.11: Visualization of the concept of a PRIMARY SEGMENT. *Top*: Two PRIMARYs are pictured in brown and purple. The purple PRIMARY's CHUNK is overlaid in teal. Three of that CHUNK's 12 SEGMENTs are shown. *Middle*: Same as the top panel except the CHUNK in teal has been removed. This more clearly shows that some of the CHUNK's SEGMENTs now extend beyond the PRIMARY's boundaries. If the upper CHUNK's SEGMENTs were visualized, a subset of its SEGMENTs would overlap those shown. *Bottom*: The SEGMENTs that extend outside their PRIMARY are cropped to create new regions called PRIMARY SEGMENTs.

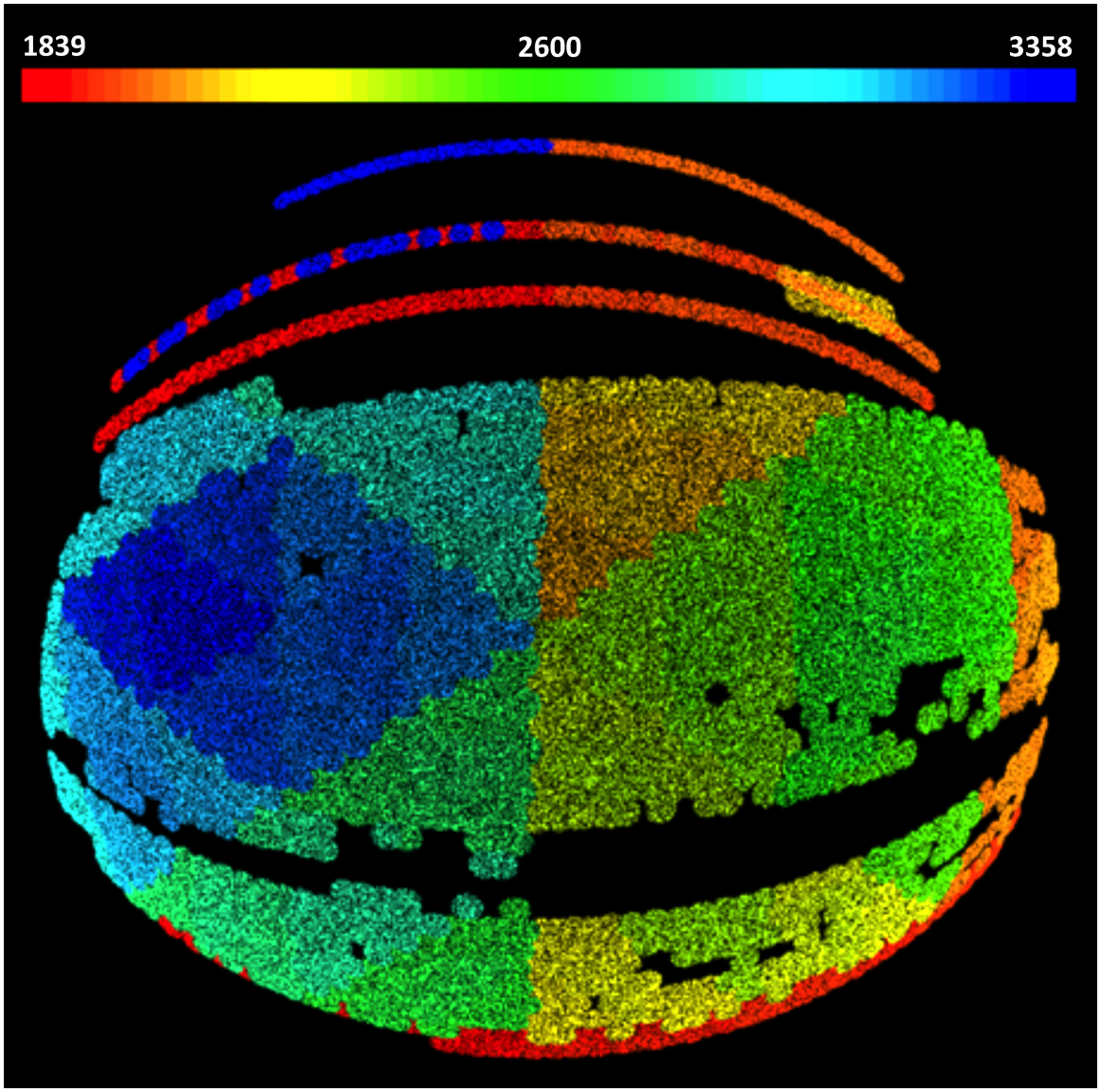

Figure 2.12: Spatial representation of the TILEs defined within the SDSS DR6 database. Each circular TILE is assigned a color based upon its *regionID*. The order of the *regionID*'s does not convey the order in which spectroscopic observations were conducted within and between data releases.

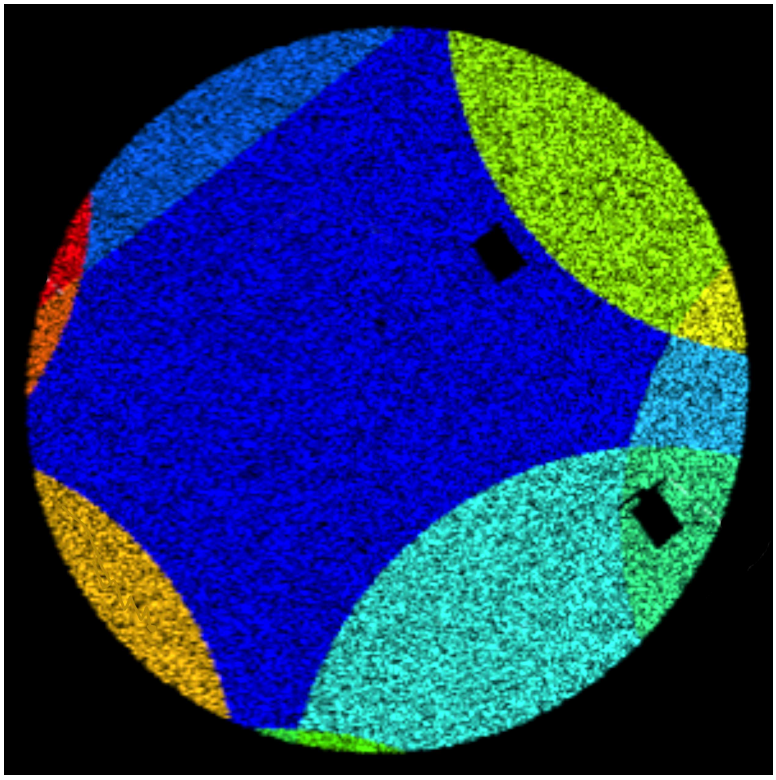

Figure 2.13: Visualization of DR6 TILE 550. This TILE is comprised of 12 SECTORs, each of which is represented by a random color. These SECTORs are created by TILE 500's intersection with six other TILEs and one great circle constraint (straight line in the upper-left). Two roughly rectangular titling masks, shown in black, reduce the areas of the two SECTORs within which they reside. The geometric description of each tiling mask is directly incorporated into the definition of its SECTOR.

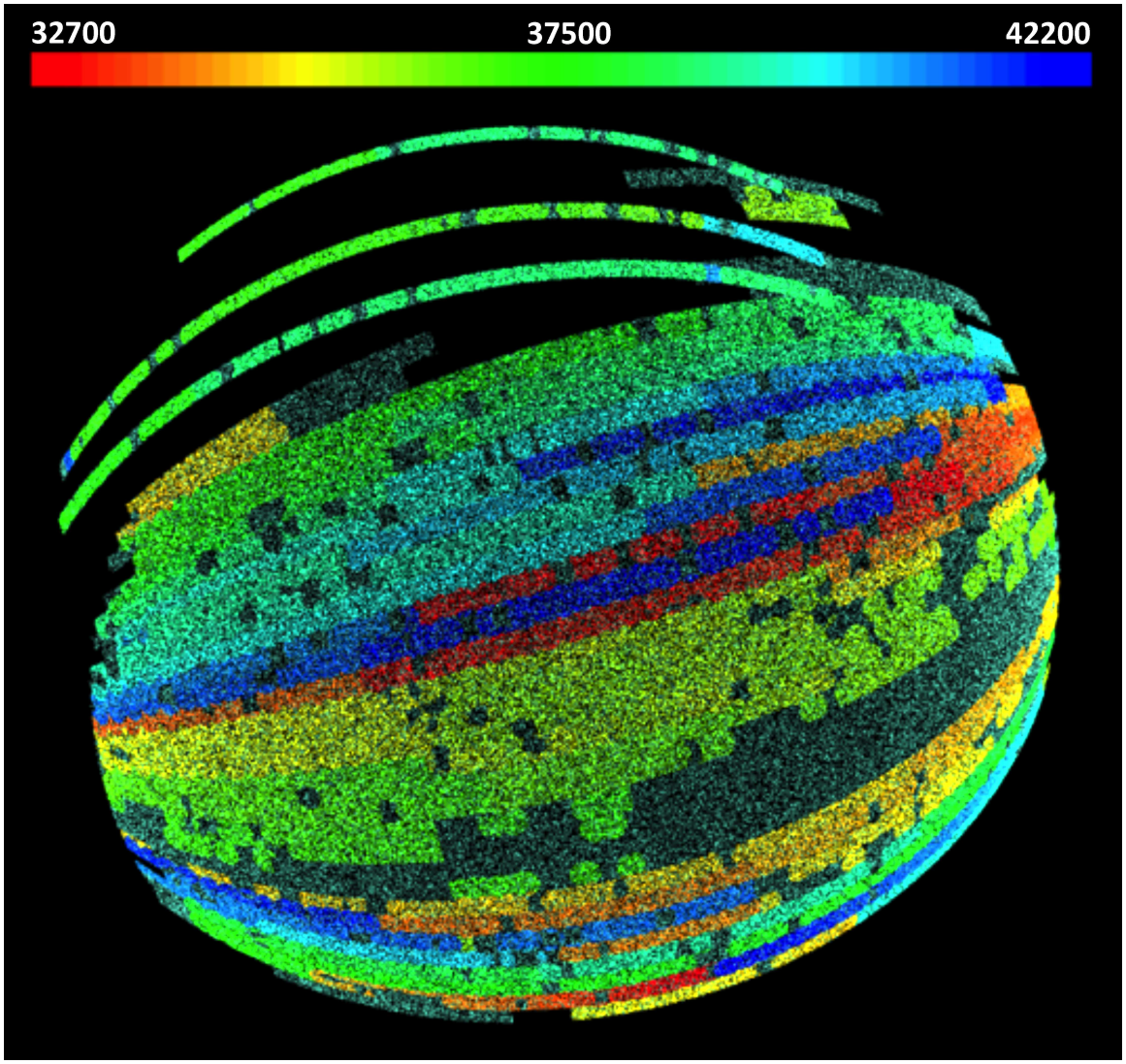

Figure 2.14: DR6 spectroscopic footprint (high density colored points) overlaid atop the photometric footprint (lower density blue-green points). Each of DR6's 9464 SECTORs has been given a unique color based on its *regionID*.

2.2.3 Region Algebra

While the SDSS uses a few different methods to describe a region's geometry, I exclusively utilize the so-called *constraint conditions*. The basic idea is that regions like SECTORs and SEGMENTs have multiple sides, each of which can be considered a constraint. Any object that satisfies all of a region's constraints must lie within it. (For more, read *There Goes the Neighborhood: Relational Algebra for Spatial Data Search*, Gray et al., 2004.)

Sloan treats each constraint as a planar intersection of the sky's unit sphere, as in Figure 2.15. The resulting small or large circle is described by four parameters: the three Cartesian components of $\hat{\boldsymbol{n}}$, the unit vector which points towards its center, and $c \equiv \cos\theta$ where θ is the circle's angular radius. This area is referred to as a *halfspace* since the plane divides three-dimensional space in half.

A point $\hat{\boldsymbol{x}}$ on the unit sphere lies within the halfspace if $\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{x}} - c > 0$, an inequality referred to as a *constraint condition* or a *halfspace constraint*. Circles with small angular radii have $c \approx 1$. With great circles, $c = 0$. Halfspaces with $c < 0$ correspond to areas greater than a hemisphere.

More complicated areas are created by intersecting multiple halfspaces. In general, these intersections are called *convexes*. For example, a SEGMENT is a convex with four halfspace constraints (i.e. four sides). A point that simultaneously satisfies all four of those constraints lies within the boundaries of the SEGMENT. Convex constraints are extracted from the SDSS database's `Region`, `RegionConvex`, `Segment`, and `HalfSpace` ta-

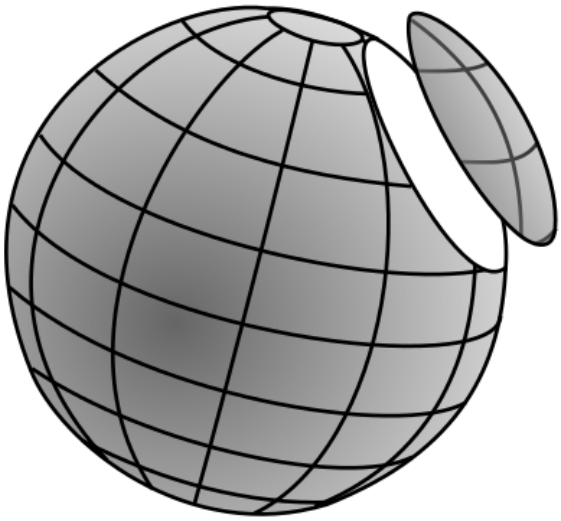

Figure 2.15: A plane intersects a sphere, producing a spherical cap. The circumference of the cap is the small circle. The SDSS relies heavily upon the intersections of small circles to describe regions. Image by Cronholm144, used under CC BY-SA 3.0.

bles.

The halfspaces and their associated inequalities comprise a *region algebra* and provide a convenient framework for determining whether or not an object occupies a region. In general, for a point $\hat{\boldsymbol{x}}$ to lie within a convex with m constraints,

$$\hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{x}} - c_i > 0 \quad \forall\, i = 1, 2, \ldots m. \tag{2.3}$$

Note that this region algebra does not require trigonometric functions, but only relatively inexpensive dot products. The increase in speed this algebra provides was indispensible for efficiently executing many of the simulations discussed in the pages to come. An example that uses four constraints to represent PRIMARY 208 is shown in Figure 2.16.

SEGMENTs, PRIMARYs, CHUNKs, and TILEs are both regions and convexes, while

SECTORs are regions formed from the union of one or more convexes. Each of a SECTOR's convexes received a *convexid* in the DR6 `HalfSpace` table that range between 0 (the 1^{st} convex) and 11 (the 12^{th} convex). A point lies within a SECTOR as long as it satisfies all of the constraint conditions of *any* of its convexes.

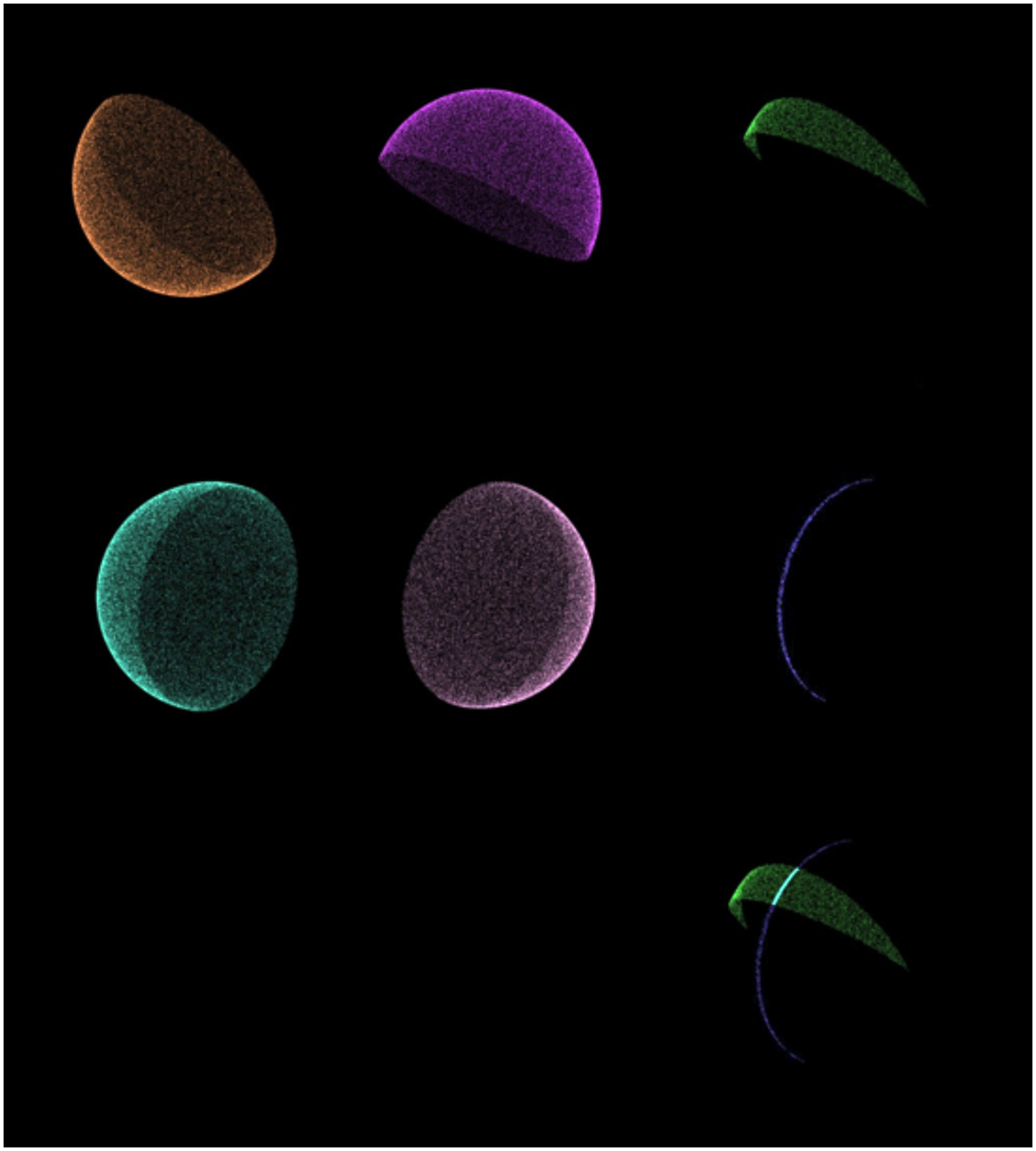

Figure 2.16: Using halfspace constraints to specify the boundaries of PRIMARY 208. Each of the shapes is formed by subjecting uniformly distributed random points on the unit sphere to one or more constraint conditions. The four hemispheres in the upper left hand corner each respectively satisfy one of the PRIMARY's four halfspace constraints. All four are formed with great circles where $c = 0$. Points in the third column lie in the intersection of the previous two. The green wedge (row 1) captures the length of a run while the thin purple strip (row 2) follows a STRIPE from pole to pole. The image in the lower right hand corner shows the union of the "wedge" and "strip" with the intersection, which represents PRIMARY 208, highlighted in cyan.

2.3 The Main Galaxy Sample

A main scientific driver of SDSS-I and SDSS-II was revealing the clustering properties of the Universe on both large and small scales. To this end, the Legacy Survey targeted two samples of galaxies — the Main Galaxy Sample (MGS) and the Luminous Red Galaxies (LRGs). The former is a brighter, flux-limited sample. The latter is redder, deeper color-limited sample (Eisenstein et al., 2001).

It is well known that luminous galaxies cluster more readily (i.e. with greater bias, which is discussed in §3.2) than less luminous ones (e.g. Davis et al., 1988; Hamilton, 1988; Park et al., 1994; Norberg et al., 2001; Zehavi et al., 2002; Verde et al., 2002) and in the interest of uniformity, it is often advantageous to consider each group separately. Because the handling of systematic noise is a major focus of this thesis, and because the effects of photometric zero-point offsets are more directly investigated with the MGS, the Main Galaxy Sample is our primary focus.

To best probe galaxy clustering on all scales, the MGS sample needed to be complete, or in other words nearly all galaxies that could be detected, should be. As such, the MGS was designed to be a full-sky, three-dimensional sample. This stands in contrast to pencil-beam surveys of the past where narrow fields of view impeded studies of clusters and measures of high frequency clustering properties (e.g. Szapudi & Szalay, 1996).

The sample also needed to be uniform. LRGs, which are intrinsically more luminous, are visible to much greater distances than MGS galaxies. All things being equal, objects at high redshift tend to be of lower apparent magnitude. Therefore the MGS needed to be

subject to some upper magnitude limit to separate the two samples.

In deference to both goals, an MGS magnitude limit was established as a compromise — bright enough to achieve uniformity but faint enough to collect a complete, high surface density of galaxies. The limit had to be minimally sensitive to reddening and variations in the sky background. Also, object detection needed to be of a significantly high signal-to-noise (S/N) ratio to avoid false positives.

Furthermore, Strauss et al. (2002) established five additional criteria for the MGS:

- Galaxies should yield an accurate selection function
- Galaxies should jointly maximize uniformity and inclusion of a broad range of physical properties
- Selection criteria should be based on physically meaningful parameters
- Selection criteria should prioritize objects that will easily yield spectra
- All things being equal, keep the selection algorithm simple

This section achieves several goals. First, it offers definitions for Main Galaxy Sample targets. Next, it splits the MGS into three mutually exclusive groups, each of which will be studied separately. Finally, the selection function for MGS targets is introduced and parameterized, and the effect of zero-point magnitude errors on the expected number of galaxies as a function of redshift is provided.

2.3.1 Selection Criteria

It was decided that for simplicity, the MGS magnitude limit would be in one bandpass only. The r-band was ultimately chosen to minimize reddening effects. The Petrosian magnitude r_P was selected over other options (e.g. fiber, de Vaucouleurs) since it is insensitive to foreground extinction, sky brightness and cosmological surface brightness dimming. Furthermore, having a faint magnitude limit in the Petrosian system did not disqualify a galaxy for simply having a low surface brightness.

The most preliminary MGS cut required that an object be 5σ above the sky after smoothing with a Point Spread Function filter (PSF). This high S/N was also useful in creating the exponential and de Vaucouleurs fits needed to assign a model magnitude. To distinguish between stars and galaxies, it is required that $r_{model} \leq r_{PSF} - 0.3$. To filter out stars, objects flagged as SATURATED in the database were rejected. So too were BRIGHT and BLENDED (unless they were children of a deblending algorithm) objects.

After these initial checks, a faint magnitude limit of $r_P \leq 17.77$ was enforced. This value was selected so that the MGS would yield approximately 90 galaxies/deg^2 . Note that throughout r_P refers to the extinction-corrected Petrosian magnitude in the r-band or (*petromag_r-extinction_r*) as stored in the database.

One caveat is that while most MGS targets were identified during early runs, photometric calibrations changed to ubercal during DR7. In some cases this led to the magnitudes of previously identified MGS targets being adjusted such that $r_P > 17.77$.

Galaxies for which $r_P \leq 17.77$ almost always have large surface brightness den-

sities, which is an important prerequisite for taking reliable spectra. The surface density was quantified through the half-light surface brightness μ_{50}. Everything for which $\mu_{50} \leq 23.0\,\mathrm{mag/arcsec^2}$ in r was considered a galaxy and had its *primTarget* flag in the database set to GALAXY. About 99% of galaxies which survived the faint Petrosian magnitude limit made this cut.

Of the remaining galaxies, 60% of those for which $23.0 < \mu_{50} \leq 24.5$ were improperly deblended objects like star diffraction spikes and spiral arms of galaxies. If the object's local background was similar to the average background, it was assumed to be isolated and admitted. All other galaxies were admitted if $r_{fiber} \leq 19$, since these objects were likely to easily yield quality spectra. There were a couple of other cuts to overly bright objects, but these only affected about 0.01% of candidates. A summary of these selection criteria is presented as a flow chart Figure 2.17.

The MGS target criteria were imperfect. Stars were sometimes confused with galaxies and vice-versa. Identifying galaxies was the primary focus of the survey, so minimizing false negatives was prioritized over admitting false positives, especially since the latter could hopefully be identified by follow-up spectroscopy.

The Main Galaxy Sample was officially completed in 2008 at the conclusion of SDSS-II. The final data release to contain new MGS objects was DR7. Since DR8, MGS data has been stored within the database as part of the Legacy Survey and identified through a flag set in the field *legacy_target1*. In principle, it is preferable to query the MGS from later data releases as they are inclusive of earlier releases and tend to have optimized photometric

	Count
objects in `PhotoPrimary`	230,417,920
objects defined as MGS objects	824,287
MGS targets with template photo-z's	824,286
MGS targets with ANN photo-z's	816,401
MGS targets with spectra	577,436
MGS targets with spectra and K corrections	572,819

Table 2.1: Census of DR6 objects identified as MGS targets through their photometric properties.

calibrations and astrometry.

Ultimately, galaxies that survived these cuts were defined to be *MGS targets* and became candidates for spectroscopic observation. Depending on which specific criteria were met, one of more of the flags listed below were set in the *primTarget* database field of the `PhotoPrimary` table:

64 = 'TARGET_GALAXY'

128 = 'TARGET_GALAXY_BIG' and/or

256 = 'TARGET_GALAXY_BRIGHT_CORE'.

For more details on how these queries were executed through SQL, reference Appendix A.

There are approximately 230 million primary photometric objects in the DR6 database but only 824,287 of those are MGS targets. In DR6 almost all objects have both template and training-set photo-z's. The number of targets with spectra is around 70% with over 99% of those having K corrections. Table 2.1 summarizes these results.

2.3.2 Pristine MGS Galaxies and Non-Pristine MGS Objects

Objects that satisfy the MGS criteria (i.e. *MGS targets*) detailed in the previous section fall into two broad categories — those with quality spectra (hereafter *MGS galaxies*) and those without (hereafter *MGS objects*). While MGS galaxies readily yield their radial distances, MGS objects complicate the creation of a complete and accurate map of the local Universe. Failure to account for MGS objects systematically underestimates the true number of MGS targets in many regions of space.

It is tempting to ignore MGS objects entirely. If the number of galaxies expected per unit redshift (in the absence of clustering) is normalized using only MGS galaxies, then the overdensities of said objects (see §3.1) might approximately equal the values obtained from using all MGS targets. If the angular distribution of MGS objects was perfectly isotropic, then such an approach would certainly be valid, at least for measuring densities.

In reality, the percentage of MGS objects is large enough to add significant variance to that approximation. Furthermore, the distribution of MGS objects is far from isotropic. Fiber collisions often serve to leave overdense areas of the sky underobserved. Edges of the survey that will be tiled and spectroscopically observed in a future data release leave MGS objects in their wake. That is, we know the lines-of-sight they lie along, but can only make clumsy guesses as to their distances.

This section takes the first step in solving this problem. We split the MGS targets

of Strauss et al. (2002) into two groups. The first exclusively contains MGS galaxies with high-quality spectra. We refer to these as *pristine galaxies*. The second contains everything else (e.g. targets without spectra, targets with low quality spectra). We refer to these as *non-pristine objects*.

Before delving into the details of these two groups, we impose one additional criterion. In order for an MGS target to enter our sample, it must satisfy a bright magnitude limit of $r_P > 15$. This is designed to both maintain the uniformity of the sample and flatten out the selection function at low redshift. This additional constraint rejects about 2.8% of otherwise eligible targets. It should be assumed that all references to MGS objects, galaxies or targets incorporate this new criterion unless otherwise stated.

2.3.2.1 Pristine Sample

To be designated a pristine galaxy, an MGS target must:

1. have a redshift (i.e. *specObjID* cannot be set to 0) that is of high-confidence and quality,
2. be spectrally classified as a galaxy (as opposed to a star, QSO, etc.) through field *specClass*,
3. reside in the redshift range $0.02 \leq z \leq 0.30$, and
4. have an absolute magnitude $M_r \leq -17$.

	Count	**Percentage**
MGS targets with $r_P > 15$	801,281	N/A
above with spectra	562,743	70.2
above with *specClass* = galaxy	551,847	98.1
above at $0.02 \leq z \leq 0.22$	526,886	95.5
above with good *zStatus*	497,682	94.5
above with *zConf* ≥ 0.9	480,569	96.6
above where $M_r < -17$	480,569	100

Table 2.2: Census of 480,569 DR6 MGS targets that satisfy the *pristine sample* criteria. Each row contains the number and percentage of objects remaining from the previous row after the filtering condition is applied. In practice, the absolute magnitude condition was redundant as all DR6 pristine galaxies that satisfied the first three conditions also satisfied the fourth.

The first two conditions maximize the probability that members of the *pristine* sample are actually galaxies. Like the apparent magnitude condition, the absolute magnitude criterion is introduced to maintain sample uniformity.

Galaxies in the low-redshift range $z < 0.02$ posed a couple problems. After spectroscopic analysis, many were discovered to be non-galaxies (usually stars) that made it through the MGS target selection pipeline. Peculiar velocities were also a concern in this regime, as a number of galaxies had redshifts close to or less than 0. If left in, these galaxies would artificially drive up measurements of the absolute magnitude and erroneously skew the selection function. The lower redshift limit of $z \geq 0.02$ was chosen to avoid the worst of these problems.

Table 2.2 shows how the number of MGS objects is reduced by cuts along the way. The median redshift of the pristine galaxies is approximately $z = 0.1$.

specClass	Identifier	Number Count
UNKNOWN	0	2
STAR	1	1
GALAXY	2	480,569
QSO	3	2427
STAR_LATE	6	1

Table 2.3: Distribution of pristine galaxies if the criterion forcing them to be of *specClass* GALAXY is lifted. Some of the possible values of *specClass* and their integer identifiers in CAS occupy the first and second columns.

2.3.2.2 Non-Pristine Sample

All MGS targets not defined to be pristine galaxies are *non-pristine objects*. Members of this subsample fall into one of three mutually exclusive groups — those

- whose quality spectroscopic data reveal that they are not actually galaxies,
- whose spectral classifications and/or redshifts are of low confidence, and
- with no or useless spectra

2.3.2.2.1 NON-GALAXIES

An MGS target's spectrum can reveal whether it is a galaxy, star or some other object. Table 2.3 reports the distribution of pristine galaxies if the $specClass = 2$ criterion is lifted. Because *specClass* is set only after spectral analysis, targets that satisfy the photometric MGS target selection algorithm can actually be misidentified stars, QSOs or other objects. The redshift distribution of these misidentified targets is plotted in Figure 2.18.

The spectral classification algorithm excludes about 0.5% of otherwise pristine galaxies. Most of the misidentified objects are quasars, which have their own host galaxies but

lie at the far end of the MGS/active galaxy continuum. It is likely that many other targets defined as GALAXY also have quasars, but with beams that are not directed towards us.

This raises the difficult question of how to handle this information. Because there is no bias in the tiling algorithm, it is reasonable to assume that approximately 0.5% of MGS objects are non-galaxies as well. When counting them in, say, redshift bins we might feel compelled to downweight the resulting counts according to the equation given by the data in Figure 2.19. Yet doing so would ignore the fact that many objects deemed *pristine galaxies* might no longer be if their random orientations with respect to Earth happened to be different.

The resolution to this question largely depends on the type of information one wishes to extract from the data. As such, we contend that there is no one "right answer". Whether to include galaxy misidentification corrections is left to reader to handle as they see fit. They will not, however, be incorporated into the galaxy counting analysis in Chapter 6.

2.3.2.2.2 OBJECTS WITH NO OR USELESS SPECTRA

About 30% of MGS targets (i.e. 238,538 objects) lack spectra as indicated by their having *specObjID* $= 0$. Another 29,478 have nonzero *specObjID*'s but have *zStatus*'s equal to 0, 1 or 2 which are, respectively, "redshift not taken", "redshift measurement failed", and "redshift cross-correlation and emz both high-confidence but inconsistent".

None of these objects possess useful spectral information therefore no redshift range or *specClass* criteria are imposed. Spectrally-defined K corrections (see Appendix C for

zStatus	Identifier	Number Count
Redshift determined from cross-correlation with low confidence	5	2386
Redshift determined from em-lines with low confidence	8	133
Redshift determined "by hand" with low confidence	10	304

Table 2.4: Census of DR6 MGS targets that enter the low-quality sample due to having their *zStatus* flag set to 5, 8, or 10. Descriptions of these flags are provided in the first column, while the number of targets that satisfy them are provided in the third column.

details) are also unavailable, so no absolute magnitude limit can be applied. If the statistics of the pristine galaxies are representative, this omission should have negligible effect.

However, these MGS objects still possess useful information including template photo-z's, training-set photo-z's and angular positions relative to other pristine galaxies. Section 6.1 exploits the correlations between these quantities and radial distance as a way to improve the counting of targets within discrete volumes. See Appendix C for the explicit queries used to extract these objects.

2.3.2.2.3 OBJECTS WITH LOW-QUALITY SPECTRA

In the middle ground between pristine galaxies and no-redshift objects are MGS objects with low-quality spectra. These include objects with *zconf* < 0.9 and any with *zStatus* flags listed in Table 2.4. As with the no-redshift group, the $M_r \leq -17$ criterion is dropped. Due to potentially compromised spectra, all redshifts and *specClass* designations are permitted. This over-inclusion of objects will be addressed statistically during the counting analysis.

There are 22,850 MGS targets in the low-quality object group. Their *zConf* distribution

is illustrated in Figure 2.20.

2.3.2.3 Spatial Distribution

The spatial distributions of pristine galaxies, no-redshift objects and low-quality redshift objects are presented in Figures 2.21, 2.22, and 2.23 respectively. Regions with higher densities of no-redshift objects are often the same as regions with no pristine galaxies. This is the result of normal SDSS observing strategy. Photometry always precedes spectroscopy, and many of the no-redshift objects will be converted to pristine galaxies in DR7.

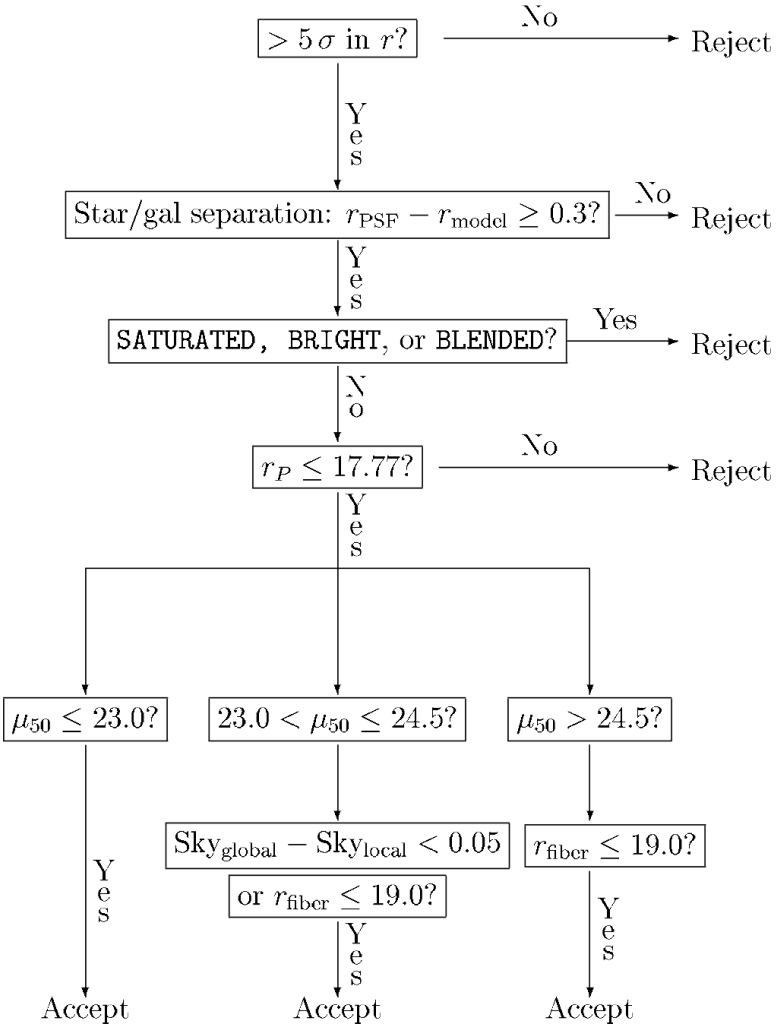

Figure 2.17: Schematic flow diagram from Strauss et al. (2002) depicting the selection algorithm for MGS targets. Explanations of each quantity in this figure are provided in the text.

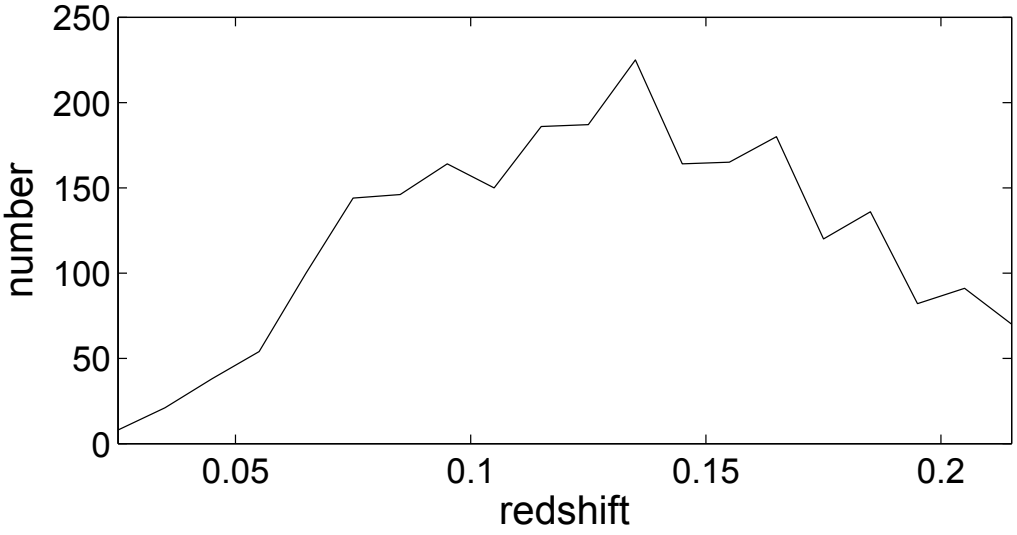

Figure 2.18: Distribution of MGS targets that satisfy all *pristine galaxy* criteria except *specClass* $= 2$. Targets are counted in redshift bins of size $\Delta z = 0.01$.

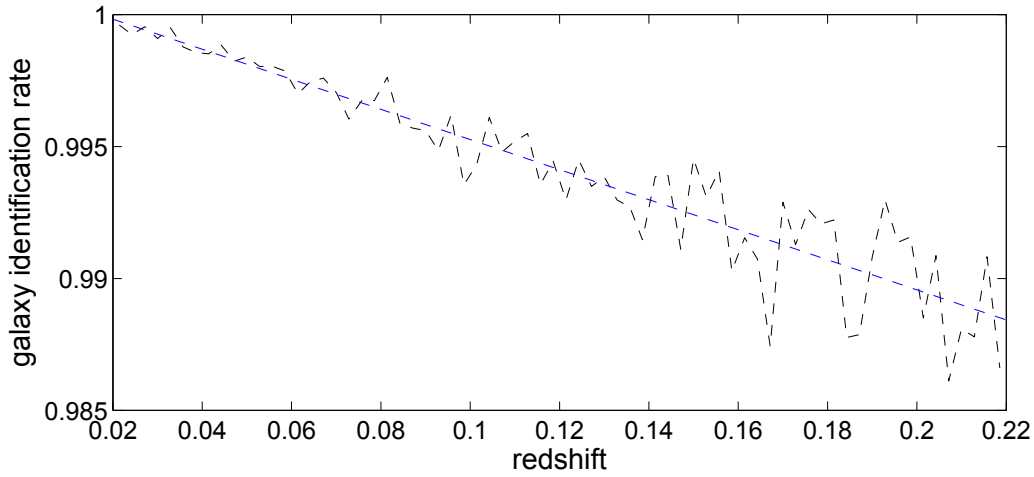

Figure 2.19: Probability that an DR6 MGS object is not a GALAXY. The vertical axis is a fraction in which the numerator is the number of pristine galaxies and the denominator is that same number plus the number stars, QSOs and other objects that would erroneously be considered pristine galaxies if not for their lack of GALAXY designations in the *specClass* field. Targets are counted in 70 redshift bins uniformly spaced between $z = 0.02$ and $z = 0.22$. A linear best-fit with equation $m(z) = -0.057z + 1.00$ is included.

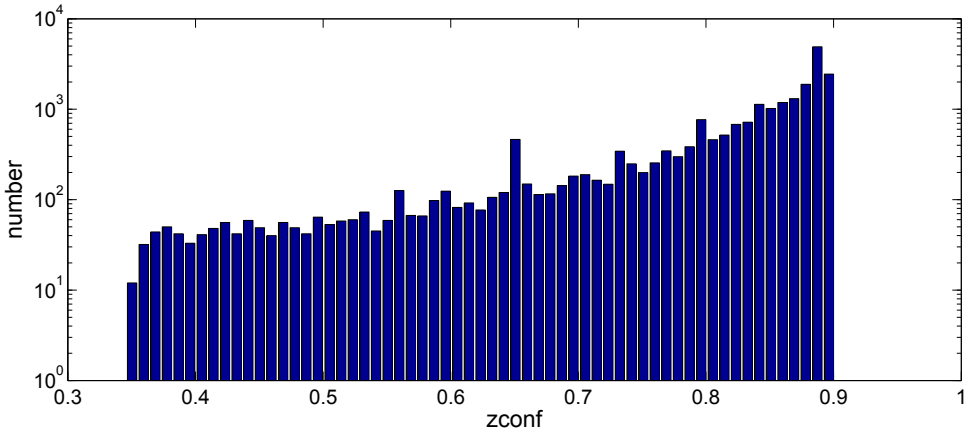

Figure 2.20: Histogram of redshift confidences for DR6 MGS objects in the low-quality group. Objects are counted in bins of width 0.09.

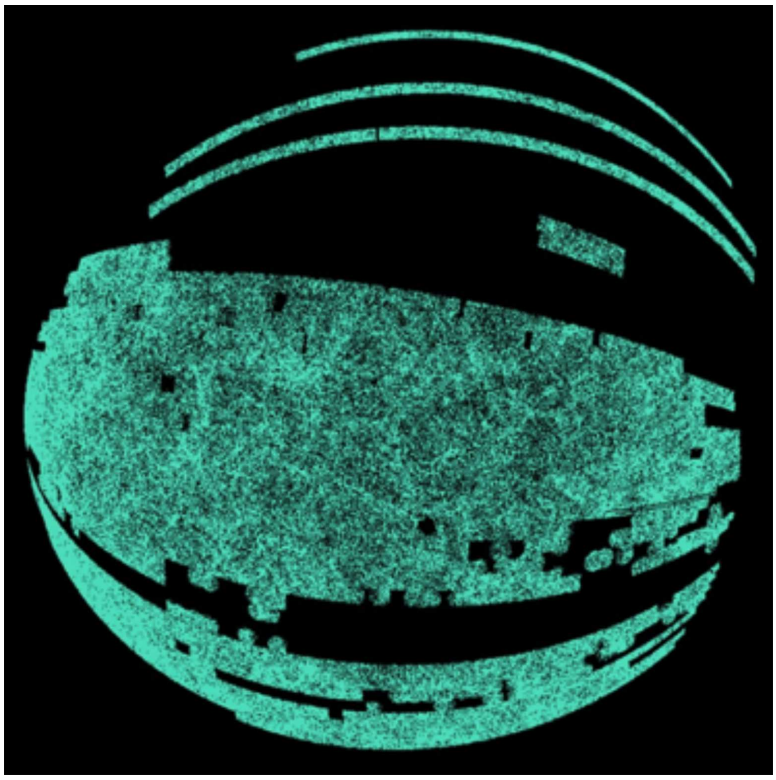

Figure 2.21: Angular distribution of DR6 MGS pristine galaxies.

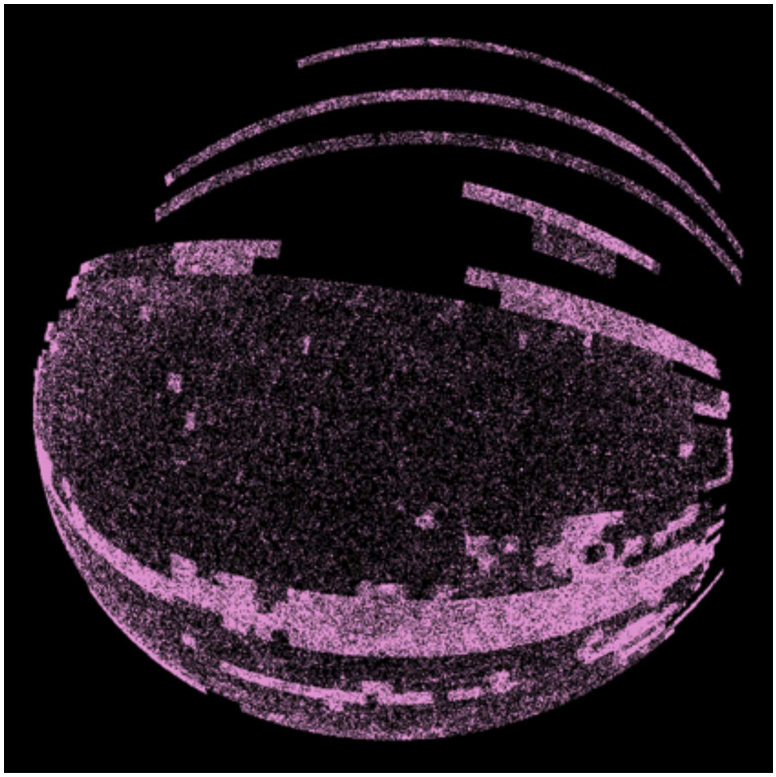

Figure 2.22: Angular distribution of DR6 MGS no-redshift objects.

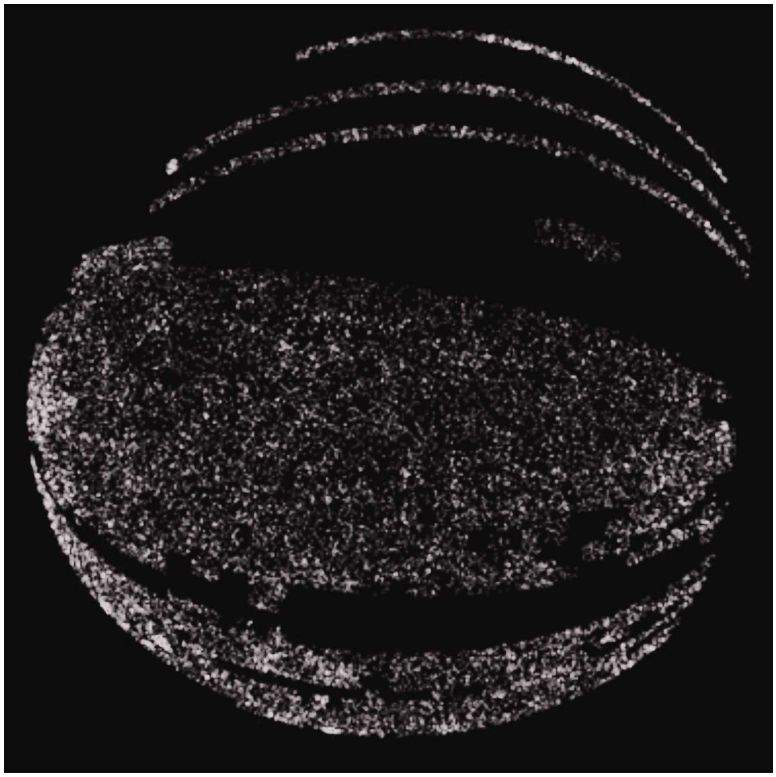

Figure 2.23: Angular distribution of DR6 MGS low-quality redshift objects.

2.3.3 Luminosity and Selection Functions

Before one can measure the overdensity of galaxies in a region of space, one must first know the number expected there in the absence of clustering. In a magnitude limited survey this is arrived at through the selection function. The selection function depends intimately upon the limiting magnitude of the survey and can be empirically parameterized by using the true distribution of galaxies.

This section provides background on the selection function as well as the luminosity function from which it derives. Optimal parameters for the selection function are determined through a maximum likelihood method. Equations that yield an expected number of galaxies per unit redshift are offered. Finally, the process of developing a model of the photometric zero-points begins by quantifying how the expected number of galaxies varies as a function of limiting magnitude.

Derivations for many of the functions referenced in this section are provided in Appendix A.

2.3.3.1 Parameterization

The luminosity function $\Phi(L)$ describes the number of galaxies in a volume dV in the luminosity range $(L, L + dL)$. A parameterized version $\Phi(L)$ was introduced by Schechter (1976) and takes the form of equation (2.4),

$$\phi(l)\, dl = \phi^* l^\alpha e^{-l}\, dl, \tag{2.4}$$

where $l \equiv L/L^* = 10^{(M^*-M)/2.5}$. The parameters are a characteristic magnitude M^*, a faint-end slope α and a normalization parameter ϕ^* with units of number density. An equivalent, and for our purposes more useful, version of equation (2.4) replaces luminosity with absolute magnitude,

$$\phi(M)\,dM = \left(-\frac{\ln(10)}{2.5}\right)\phi^*\left(10^{(M^*-M)/2.5}\right)^{\alpha+1}\exp\left[-\left(10^{(M^*-M)/2.5}\right)\right]dM. \qquad (2.5)$$

Parameterizations of the luminosity function have been performed by the likes of Sandage et al. (1979), Efstathiou et al. (1988), and Blanton et al. (2001). Their goal was to parameterize $\phi(M)$ by minimizing a likelihood function involving $p\,(M_j|z_j)$, or the conditional probability of observing a galaxy with an absolute magnitude M_j given its redshift z_j.

Our goal is somewhat different. We are not interested in the luminosity function in and of itself, but rather seek to understand the effect of photometric zero-points on expected galaxy counts. As such, our likelihood function involves minimizing over $p(z_j|P)$, or the probability of observing a galaxy at a redshift z_j given a set of Schechter parameters P.

While $\phi(M)$ is approximately constant in the local Universe, galaxy evolution over cosmic time suggests that it varies with z (see e.g. Sawicki & Thompson, 2006; Ryan et al., 2007). This is mainly because sub-L^* galaxies are hosted by dark matter halos from a steeper region of the dark matter mass function (Khochfar et al., 2007). It has been known for a couple decades that the relative magnitude of this steepening is small locally ($z \sim 0$)

and more significant at intermediate redshifts $z \sim 0.75$ (Lilly et al., 1995; Ellis et al., 1996). This is admittedly larger than the MGS median redshift of $z_0 \cong 0.1$, but because the effect is more prominent for late-type galaxies like those in the MGS, it warrants consideration.

This redshift dependent evolution can be modeled through the more general *evolving Schechter luminosity function*. It is identical to the original except that M^* is permitted to evolve with redshift through a fourth parameter B,

$$M^* \rightarrow M^* - 2.5 \log \left(\frac{1+z}{1+z_0} \right)^B . \tag{2.6}$$

Parameterized over a large enough volume, $\phi(M)$ is thought to be independent of one's choice of origin. However, there is a difference between the distribution of galaxies that are *present* and the distribution of galaxies we *actually detect*. The probability of detecting galaxies locally is approximately 1, but within any magnitude limited survey this value will decrease monotonically with z. In principle, there is some characteristic redshift beyond which all galaxies of a particular type are too dim to detect.

The probability distribution that a galaxy *actually present* at redshift z is *observed* is given by the radial *selection function* $S(z)$,

$$S(z) \propto \int_{M_{min}(z)}^{M_{max}(z)} \phi\left(M, z\right) \, dM. \tag{2.7}$$

$\phi\left(M, z\right)$ is integrated over the possible range of absolute magnitudes at z where

$$M_{max}(z) = \min\left(M_{max}(z)', M_{max}\right), \tag{2.8}$$

$$M_{min}(z) = \min\left(M_{min}(z)', M_{min}\right), \tag{2.9}$$

and

$$\begin{Bmatrix} M_{max}(z)' \\ M_{min}(z)' \end{Bmatrix} = \begin{Bmatrix} m_{max} \\ m_{min} \end{Bmatrix} - (5\log d_L(z) + 25) - k(z). \tag{2.10}$$

As discussed, the absolute magnitude is capped at $M_{max} = -17$. Both to ensure the available magnitude difference $M_{max}(z)' - M_{min}(z)'$ remains constant for all z and to render the maximum likelihood estimator that will parameterize the Schechter function analytically solvable, the minimum absolute magnitude is set to $M_{min} = -19.77$.

The integral over $\phi\left(M, z\right)$ is more directly evaluated using the formulation of equation (2.4). Ignoring ϕ^*,

$$\int_{M_{min}(z)}^{M_{max}(z)} \phi\left(M, z\right)\, dM = \int_{l_{min}(z)}^{\infty} l^{\alpha} e^{-l}\, dl - \int_{l_{max}(z)}^{\infty} l^{\alpha} e^{-l}\, dl, \tag{2.11}$$

where

$$\begin{Bmatrix} l_{max}(z) \\ l_{min}(z) \end{Bmatrix} = 10^{\left(M^* - 2.5 \log\left(\frac{1+z}{1+z_0}\right)^B - \begin{Bmatrix} M_{min}(z) \\ M_{max}(z) \end{Bmatrix} \right)/2.5} \tag{2.12}$$

The integrals from equation (2.11) possess the form of an upper incomplete gamma function, such that

$$\int_{M_{min}(z)}^{M_{max}(z)} \phi\left(M, z\right)\, dM = \Gamma\left(\alpha + 1, l_{min}(z)\right) - \Gamma\left(\alpha + 1, l_{max}(z)\right). \tag{2.13}$$

The probability of an observed galaxy being drawn from the redshift range $(z, z + dz)$ is

$$p_{exp}(z)\, dz = \frac{1}{(d_H \int d\Omega) c} S(z) \left(d_H \frac{\chi(z)^2}{E(z)} dz \int d\Omega \right), \tag{2.14}$$

where the normalization factor c equals

$$c = \int_{z_{min}}^{z_{max}} S(z) \frac{\chi(z)^2}{E(z)} dz. \tag{2.15}$$

The number of expected galaxies within that range is

$$n_{exp}(z)\, dz = N p_{exp}(z)\, dz, \tag{2.16}$$

where N is the total number of galaxies in the survey. Various normalizations of equation (2.16) are possible, but we let N refer to all MGS targets within the DR6 improved spectroscopic footprint.

If we assume the redshifts are independent and free of any clustering signal, the likelihood that a set of galaxies with redshifts $\{z_1, \ldots, z_N\}$ is drawn from a distribution with parameters $\{\alpha, M^*, B\}$ is

$$\mathcal{L} = p\left(z_1, \ldots, z_N | \alpha, M^*, B\right) = \prod_{i=1}^{N} p_{exp}\left(z_i | \alpha, M^*, B\right). \tag{2.17}$$

The log likelihood function to be maximized is

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln p_{exp}(z_i) = \sum_{i=1}^{N} \left[\ln S(z_i) - \ln c + \ln \frac{\chi(z_i)^2}{E(z_i)}\right]. \tag{2.18}$$

The latter term in the log likelihood function is constant with respect to $\{\alpha, M^*, B\}$ and is ignored. The problem reduces to minimizing equation (2.19) over $\{\alpha, M^*, B\}$,

$$-\ln \mathcal{L} = N \ln c - \sum_{i=1}^{N} \ln S(z_i), \tag{2.19}$$

The optimal set of parameters is found using a downhill simplex method. We found that they are $\{\alpha, M^*, B\} = \{-1.16, -21.74, 0.00011\}$. Inserted into equation (2.7), these parameters exhibit excellent agreement with the actual distribution of MGS galaxies as shown in Figure 2.24. The evolution parameter B is very nearly zero, a result that given our a priori knowledge was not wholly unexpected.

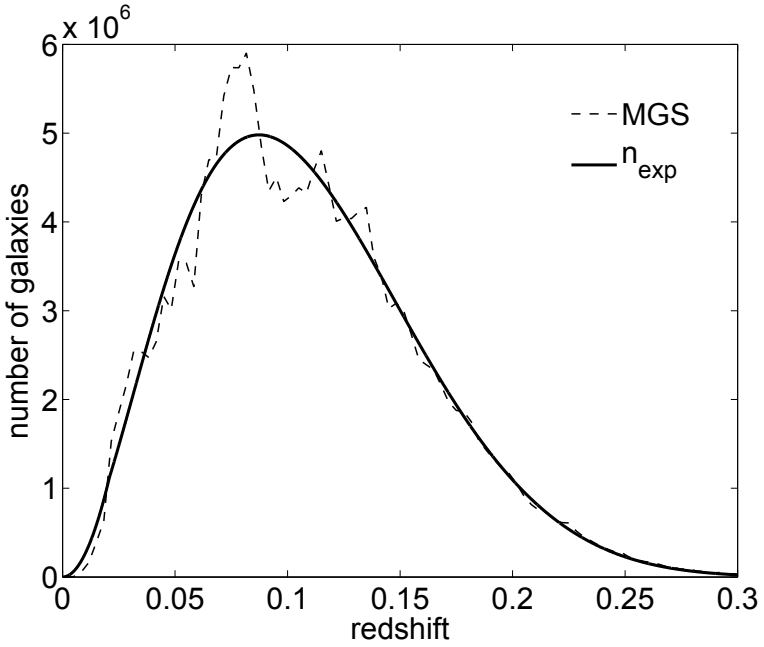

Figure 2.24: Number of expected MGS galaxies as a function of redshift. The dotted line follows the histogram of MGS galaxies from the DR8 Legacy Survey. The solid curve is the normalized number of galaxies expected in the absence of cosmological structure. It is derived from the Schechter luminosity function when $\alpha = -1.16$, $M^* = -21.74$, and $B = 0.00011$.

2.3.3.2 Effect of Limiting Magnitude on Expected Number of Galaxies

We next consider how $n_{exp}(z)$ would change if the MGS limiting magnitude shifted by a small amount to $m_{lim} + \Delta m$. A change of this kind does not alter the cosmological volume element, but it will impact the selection function $S(z)$.

Let N equal the number of MGS galaxies observed in some PRIMARY SEGMENT and let N_0 equal the number observed in the same region when $\Delta m = 0$. Then,

$$N = N_0 + \Delta N = N_0(1 + f(z)\Delta m), \tag{2.20}$$

where $f(z)$ is the percentage change in the number of observed galaxies per unit limiting magnitude. The probability of observing a galaxy that is physically present is given by the selection function. Therefore,

$$f(z) = \frac{d \ln S(z)}{dm_{lim}} \approxeq \frac{1}{S(z)} \frac{S(z)_+ - S(z)}{\Delta m}, \tag{2.21}$$

where $S(z)_+$ is the selection function for which the limiting apparent magnitude is $m_{lim} + \Delta m$. By extension, the absolute magnitude limits in $S(z)_+$ are also shifted by Δm. We assume that for both cases the MGS galaxies are drawn from roughly the same underlying distribution and therefore share the same Schechter parameters.

The choice of $\Delta m \ll 1$ matters at less than the percent level. Using the ubercalibrations we concluded $\sigma_m = 0.01$, so to avoid modeling the instantaneous change too closely, we set $\Delta m = 2\sigma_m$. As shown in Figure 2.25, the zero-points' effect on overdensities is a rapidly increasing function of redshift.

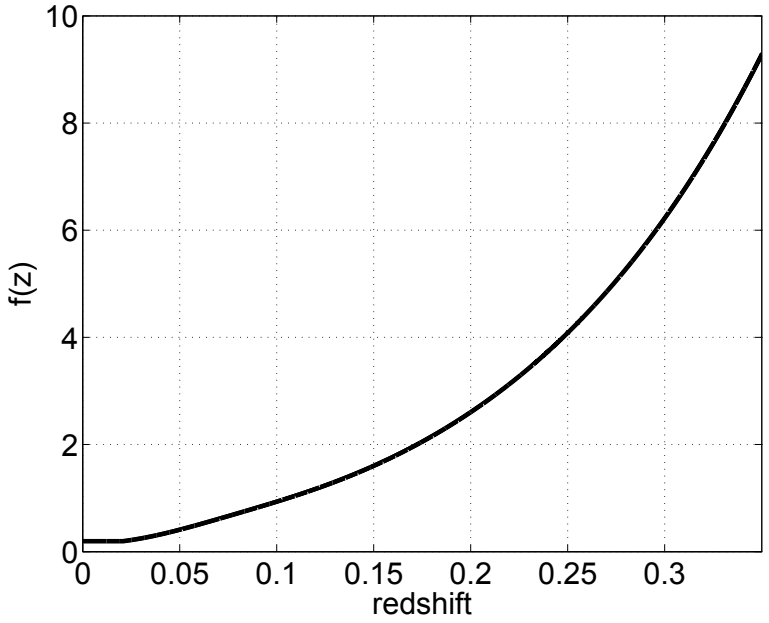

Figure 2.25: The fractional change in the number of observed MGS galaxies per unit limiting magnitude. Curve is derived from the change in the selection function of DR8 MGS galaxies near the magnitude limit $r_P = 17.77$. Zero-point offsets have the greatest impact on galaxy counts at large redshifts.

Chapter 3

Large Scale Structure

Part of our motivation for studying the positions of galaxies is to determine the *large scale structure* of the Universe. On the largest scales, cosmologists tend towards the conclusion that the Universe is both homogeneous and isotropic. This is frequently referred to as the *cosmological principle*.

However on smaller scales, the Universe presents a rich tapestry of structure. The densest regions of the Universe are superclusters of galaxies. The superclusters often act as nodes connecting long filaments of galaxies. Vast empty "bubbles" of space, or voids, fill the remainder of space.

Exactly how the Universe evolved from the Big Bang to today is very much an active area of study. One of our best tools for linking the early Universe to our present one is the study of the distribution of galaxies.

In brief, the prevailing theory is that moments after the Big Bang, quantum fluctuations

perturbed the distribution of mass away from pure homogeneity. Cosmic inflation superluminally increased the size of space and those fluctuations grew in amplitude (see Guth, 1981). Slightly overdense regions of the Universe gravitationally attracted more matter and grew bigger still. Underdense volumes became more vacuous.

As gravity pulled dense regions together, outward pressure pushed them back out. Oscillations driven by the interplay of gravity and pressure persisted until the Universe became large enough to be transparent to photons. At that point, the conditions of the Universe "froze" into place and evolved by different processes thereafter. One signature of this "freezing" are features in the power spectrum called the *baryon acoustic oscillations* (BAOs).

Perturbations from homogeneity are quantified through a spatially dependent density contrast, or overdensity, term $\delta(\boldsymbol{x})$. When these are small, they evolve in time according to *linear theory*,

$$\delta(\boldsymbol{x}, t) = \frac{g(t)}{g(t_i)}\delta(\boldsymbol{x}, t_i), \tag{3.1}$$

where $g(t)$ is a time-dependent growth factor.

Linear theory holds in the early Universe and at large scales, where gravitational forces dominate over non-gravitational ones. Here, the densities at some time t are scaled by a common factor from those at some earlier t_i for all $\boldsymbol{x}$. When sections of the Universe become particularly dense or rarified, the evolution of structure often enters a non-linear regime. During non-linear evolution overdense and underdense regions no longer scale in

the same way, and no common scaling factor exists for all space. The rank order of densities measured in discretized regions of space tend to remain the same in the non-linear regime — a fact that provides physical justification for a statistic referred to as the Gaussianized power spectrum (more on this later).

Many leading theories assume the initial distribution of overdensities follows a multivariate Gaussian distribution. Such a distribution is a natural consequence of inflationary theory (Guth & Pi, 1982; Hawking, 1982; Starobinsky, 1982; Bardeen et al., 1983). Even if this theory is incorrect, the central limit theorem reveals that the average of independent random variables drawn from different distributions tends towards the Gaussian, which in a sense makes it a safe bet. Analytically, multivariate Gaussians are relatively simple to work with and offer an attractive option for modeling. For these reasons, the work presented in this thesis assumes that galaxy overdensities follow the multivariate Gaussian distribution.

Ultimately, the distribution of galaxies we observe today is the result of the initial conditions paired with the physical evolution of the Universe. This evolution is a function of the Universe's shape, age, and composition. Therefore, measuring the distribution to a high degree of accuracy helps to constrain a number of cosmological parameters and validate the laws of physics governing the Universe's evolution. It informs us about the laws of gravity, the mass of the neutrino, and the behavior of dark energy. It constrains the conditions of the early Universe, provides information about the abundance of dark matter, and much more.

Galactic overdensities are impossible to quantify on arbitrarily small scales (e.g. less than the limiting resolution of the telescope, smaller than the characteristic size of a galaxy).

Instead, their measurements are effectively convolved with a Gaussian filter to turn galaxy counts from a set of discrete points to a smoothed field. This washes out small scale effects making certain phenomena difficult to measure, such as the scale-dependent bias between observable galaxies and the dark matter halo field, and redshift distortions due to peculiar velocities (Neyrinck et al., 2009). Between the linear and non-linear regimes lie translinear scales — those that are roughly larger than characteristic halo sizes, but too small for linear theory to apply robustly.

The research presented in these pages uses a filtering process that largely obscures the details of small scale effects. Our strongest results are for $k \lesssim 0.1h^{-1}\mathrm{Mpc}$, where k is wavenumber. This falls squarely in the linear regime. These scales are also where the photometric zero-points have their largest effect, so there is little risk of missing out on the benefits of the noise minimization technique we introduce in Chapter 7.

In this chapter, we lay out the mathematical foundations for simulating the presence of MGS galaxies and quantifying their distribution. We define overdensities more precisely and explain how they connect to the two-point correlation function. We provide background on the power spectrum and how it connects to the correlation function via a Fourier transform. We present a fiducial matter power spectrum and alter it to represent MGS galaxies through a bias factor. Finally we discuss how to modify the correlation function to account for redshift-space distortions.

3.1 The Galaxy Correlation Function

Consider the number density of galaxies $n(\boldsymbol{x}, t)$ as a function of comoving position $\boldsymbol{x}$ and time t. Let $n(t)$ (with no spatial dependence) represent the background density that would exist if the Universe were perfectly homogeneous. Then,

$$n(\boldsymbol{x}, t) = n(t)\left[1 + \delta(\boldsymbol{x}, t)\right], \tag{3.2}$$

and thus

$$\delta(\boldsymbol{x}, t) = \frac{n(\boldsymbol{x}, t)}{n(t)} - 1, \tag{3.3}$$

where $\delta(\boldsymbol{x}, t)$ is a dimensionless perturbation referred to as an *overdensity*. Overdensities are assumed to be tiny for most of the age of the Universe. Because they are fluctuations from the mean, $\langle\delta\rangle \equiv 0$. It follows that the expected number density of objects is

$$\langle n(\boldsymbol{x}, t)\rangle_{\boldsymbol{x}} = n(t)\left[1 + \langle\delta(\boldsymbol{x}, t)\rangle\right] = n(t). \tag{3.4}$$

We keep in mind that the overdensity has a time dependence and drop the t for notational simplicity.

The averaging here is a little subtle. Ideally, the average would be taken over all volumes in an infinite Universe. However, one could consider averaging one volume over multiple realizations of universes sharing the same underlying statistical properties. In

cosmology we invoke the cosmological principle to make what's known as the *ergodic assumption* — that the average over all space is the same as the average over many realizations of a single section of that space.

Next, we consider the distribution of galaxy pairs. The joint probability $dP(\boldsymbol{x}_{12})$ of finding a galaxy in each of two separate volume elements dV_1 and dV_2 separated by a vector $\boldsymbol{x}_{12}$ is

$$dP(\boldsymbol{x}_{12}) = n^2(1 + \xi(\boldsymbol{x}_{12}))\; dV_1\, dV_2. \tag{3.5}$$

Equation (3.5) serves as the definition for the *two-point correlation function* (2PCF), $\xi(\boldsymbol{x}_{12}) \equiv \xi(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ (Peebles, 1993). When $\xi(\boldsymbol{x}_{12}) = 0$, the galaxies' positions are independent of one another and $dP(\boldsymbol{x}_{12})$ reduces to a product of individual probabilities.

Following the treatment of Peebles (1973, 1980) further, we model the positions of galaxies as independent Poisson point processes modulated by fluctuations in the underlying density field. These positions possess the characteristics of a Poisson process — a large number of sampled volumes of space with a small probability that a galaxy lies within any one of them.

This suggests an alternative representation of the expected joint probability,

$$dP(\boldsymbol{x}_{12}) = \langle n(\boldsymbol{x}_1) n(\boldsymbol{x}_2) \rangle\, dV_1\, dV_2. \tag{3.6}$$

Combining equations (3.5) and (3.6) with the continued understanding that $\langle n(\boldsymbol{x}_i) \rangle = n$,

$$n^2(1+\xi(\boldsymbol{x}_{12})) = \langle n(\boldsymbol{x}_1)n(\boldsymbol{x}_2)\rangle,$$

$$\begin{aligned} n^2\xi(\boldsymbol{x}_{12}) &= \langle n(\boldsymbol{x}_1)n(\boldsymbol{x}_2)\rangle - n^2 \\ &= \langle n(\boldsymbol{x}_1)n(\boldsymbol{x}_2)\rangle - n^2 - n^2 + n^2 \\ &= \langle n(\boldsymbol{x}_1)n(\boldsymbol{x}_2)\rangle - n\langle n(\boldsymbol{x}_2)\rangle - n\langle n(\boldsymbol{x}_1)\rangle + n^2 \\ &= \langle\!\langle (n(\boldsymbol{x}_1)-n)(n(\boldsymbol{x}_2)-n)\rangle. \end{aligned} \tag{3.7}$$

We find that the correlation function amounts to a joint expectation value of the overdensities,

$$\xi(\boldsymbol{x}_1-\boldsymbol{x}_2) = \langle\left(\frac{n(\boldsymbol{x}_1)-n}{n}\right)\left(\frac{n(\boldsymbol{x}_2)-n}{n}\right)\rangle = \langle\delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2)\rangle. \tag{3.8}$$

The two-point correlation function (sometimes referred to as the autocorrelation function) is a familiar quantity across a range of disciplines. Its evaluation requires knowing only the number and expected number of galaxies in a region of space. By virtue of the cosmological principle, the Universe is isotropic. Therefore, with respect to galaxy correlations, $\xi(\boldsymbol{x}_{12})$ depends only on the distance between two points and not on direction. In theory there are an infinite number of points in the Universe separated by every distance r, which aids in averaging. In practice, the finite size of galaxy surveys and our preferred method of dividing space into discrete volumes for the purposes of counting render the number of separations countable.

Because $\langle\delta\rangle \equiv 0$, $\langle\delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2)\rangle$ is actually a covariance function. It can be a source of confusion that $\xi(\boldsymbol{x}_{12})$ is a statistical covariance function but is called a correlation function. We urge the reader to remember that "correlation function" in the context of mass and galaxy clustering has a different meaning than it does in the pure statistical sense. When referring to $\xi(\boldsymbol{x}_{12})$, we may use either adjective to describe the function depending on which aspect of its character we wish to emphasize.

The scalar field form of the correlation function in equation (3.8) can also be expressed as a real, symmetric covariance matrix in which $\Sigma_{ij} \equiv \xi(\boldsymbol{x}_i - \boldsymbol{x}_j)$ (Peebles, 1980; Vogeley & Szalay, 1996). Vogeley and Szalay represent the matrix as having three elements corresponding to signal, shot noise, and extra variance due other sources of noise,

$$\Sigma_{ij} = n_i n_j \xi_{ij} + \delta_{ij} n_i + \epsilon_{ij}. \tag{3.9}$$

Here n_i is defined to be the expected number of objects at $\boldsymbol{x}_i$, and ϵ_{ij} is other correlated noise. The second term, which introduces no cross-correlations by virtue of the Dirac delta function δ_{ij}, represents shot noise. Shot noise results from Poissonian perturbations in the number of galaxies present in a region independent of clustering.

There is another form of the correlation matrix which possesses a different weighting,

$$\Sigma_{ij} = \xi_{ij} + \frac{\delta_{ij}}{n_i} + \frac{\epsilon_{ij}}{n_i n_j}. \tag{3.10}$$

Equation (3.9) is a measure of the density and is weighted more heavily where the selection

function is high. Equation (3.10) divides by the expected number and is a measure of the overdensities. It spans a much greater volume and contains a shot noise term which dominates as the selection approaches zero. Both equations are applicable, but since our analysis primarily concerns overdensities, the latter will be employed. Both equations assume volumes i and j do not overlap.

At this point, we pause to appreciate that equation (3.8) relates the overdensities in localized regions of space to the correlation function, which is a quantification of the large scale structure of the Universe. It turns out that for large scale structure, the correlation function and its Fourier transform (the power spectrum) are very useful and important quantities quantities that provide a link between the structure of the very early Universe and today's. In the next section, we examine this same problem in Fourier space.

3.2 The Power Spectrum

3.2.1 Derivation

We now turn our attention to the frequency spectrum of density fluctuations as determined through the Fourier transform. Expanding the overdensities into a superposition of plane waves,

$$\begin{aligned}\delta(\boldsymbol{x}) &= \frac{1}{(2\pi)^{3/2}} \int d^3k\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \delta(\boldsymbol{k}), \\ \delta(\boldsymbol{k}) &= \frac{1}{(2\pi)^{3/2}} \int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \delta(\boldsymbol{x}).\end{aligned} \tag{3.11}$$

Three-dimensional Fourier transforms of this sort are always accompanied by a factor of $(2\pi)^3$, though there is no universal conventional regarding where it should be placed. Our approach is to split the factor over each transform.

The Fourier amplitude $\delta(\boldsymbol{k})$ is complex in general. It may be written in rectangular or polar coordinates, $\delta(\boldsymbol{k}) = x_k + y_k i = a_k e^{i\varphi_k}$. The fact that $\delta(\boldsymbol{x})$ is strictly real places constraints on $\delta(\boldsymbol{k})$. Because $\delta(\boldsymbol{x}) = \delta^*(\boldsymbol{x})$ (i.e. the real function equals its complex conjugate),

$$\int d^3k\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \delta(\boldsymbol{k}) = \int d^3k'\, e^{-i\boldsymbol{k}'\cdot\boldsymbol{x}} \delta^*(\boldsymbol{k}'). \tag{3.12}$$

Making the variable transformation $\boldsymbol{k}' = -\boldsymbol{k}$,

$$\int d^3k\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\delta(\boldsymbol{k}) = -\int d^3k\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\delta^*(-\boldsymbol{k}). \tag{3.13}$$

By comparison, $-\delta(\boldsymbol{k}) = \delta^*(-\boldsymbol{k})$. This is sometimes called the Hermiticity of $\delta(\boldsymbol{k})$. It shows that if you specify the Fourier modes in one half of $\boldsymbol{k}$-space, the modes in the other half are given. This observation will be applied in future sections to construct computationally efficient Fourier transform routines.

One's choice of origin in configuration-space (i.e. $\boldsymbol{x}$-space) is arbitrary by virtue of the cosmological principle. This is known as *translational invariance*. Mathematically it allows us to claim $\langle\tilde{\delta}(\boldsymbol{k}_1)\tilde{\delta}^*(\boldsymbol{k}_2)\rangle = \langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle$, where we use the notation $\tilde{\delta}$ to represent Fourier modes with an origin shifted by some vector $\boldsymbol{x}_0$. We find that shifting the origin in configuration-space introduces a phase shift in Fourier space

$$\tilde{\delta}(\boldsymbol{k}) = \int d^3x\, e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}_0)}\delta(\boldsymbol{x}) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}_0}\left(\int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\delta(\boldsymbol{x})\right) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}_0}\delta(\boldsymbol{k})\,. \tag{3.14}$$

Averaging over realizations in $\boldsymbol{k}$-space,

$$\langle\tilde{\delta}(\boldsymbol{k}_1)\tilde{\delta}^*(\boldsymbol{k}_2)\rangle = \langle\delta(\boldsymbol{k}_1)e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}_0}\delta^*(\boldsymbol{k}_2)e^{-i\boldsymbol{k}_2\cdot\boldsymbol{x}_0}\rangle. \tag{3.15}$$

The phase factors are pulled from the expectation value since they are the same for all

realizations.

$$\langle\tilde{\delta}(\boldsymbol{k}_1)\tilde{\delta}^*(\boldsymbol{k}_2)\rangle = e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}_0}\langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle = \langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle. \tag{3.16}$$

For $\boldsymbol{x}_0 \neq 0$, equation (3.16) holds only if $\boldsymbol{k}_1 = \boldsymbol{k}_2$ or $\langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle = 0$. Therefore,

$$\langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle = \langle|\delta(\boldsymbol{k}_1)|^2\rangle\,\delta^{(3)}(\boldsymbol{k}_1 - \boldsymbol{k}_2)\,. \tag{3.17}$$

Equation (3.17) reveals an important conclusion: in a Universe with translational invariance, Fourier amplitudes are independent of one another and uncorrelated in $\boldsymbol{k}$-space. However, some operations, like the noise minimization techniques presented in Chapter 7, can introduce correlations between modes.

In configuration-space, $\langle\delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2)\rangle$ is the correlation function. In Fourier space, that product is called the power spectrum $P(\boldsymbol{k})$,

$$\langle\delta(\boldsymbol{k}_1)\delta^*(\boldsymbol{k}_2)\rangle = (2\pi)^3 P(\boldsymbol{k}_1)\,\delta^{(3)}(\boldsymbol{k}_1 - \boldsymbol{k}_2)\,. \tag{3.18}$$

Furthermore, rotational invariance requires that the power spectrum depend only on the magnitude of $\boldsymbol{k}$, not its direction, such that $P(\boldsymbol{k}) = P(k)$. Since $\delta(\boldsymbol{k})$ may be written as $a_k e^{i\varphi_k}$,

$$P(k) = \langle\delta(k)\delta^*(k)\rangle = \langle a_k e^{i\varphi_k} a_k^* e^{-i\varphi_k}\rangle = \langle a_k^2\rangle. \tag{3.19}$$

This reveals that the power spectrum does not depend on the phases of the Fourier amplitudes.

3.2.2 Calculation

A standard way to evaluate density modes is by expanding each galaxy as a superposition of plane waves,

$$\tilde{\delta}(\boldsymbol{k}) = \frac{1}{\sum_j w(\boldsymbol{x}_j)} \sum_j w(\boldsymbol{x}_j) e^{i\boldsymbol{k}\cdot\boldsymbol{x}_j} - \tilde{W}(\boldsymbol{k}), \tag{3.20}$$

where $w(\boldsymbol{x}_j)$ is the weight given to galaxy j located at position $\boldsymbol{x}_j$ and $\tilde{W}(\boldsymbol{k})$ is the contribution of the survey window. Several forms of the optimal weighting have been proposed including that of Feldman et al. (1994) (hereafter FKP),

$$w(r) = \frac{1}{1 + S(r)P(k)}. \tag{3.21}$$

Like most others, this weighting is inversely related to the selection function S. Upweighting distant galaxies accounts for the fact that in magnitude-limited surveys fewer galaxies are observed at large r than are actually present. (For more background on how these equations are arrived at, see Appendix D.)

However, there are several reasons why the density expansion of equation (3.20) is non-optimal. As result of their limited angular coverage most survey volumes, including that of the MGS, are strongly anisotropic which guarantees that the Fourier modes will not con-

stitute an orthonormal basis. This introduces correlations in power between k-modes and the independence arrived at in equation (3.19) breaks down. Furthermore, the anisotropy reintroduces the directional dependence of $\boldsymbol{k}$ such that $P(\boldsymbol{k}) \neq P(k)$. When averaging power over shells in $\boldsymbol{k}$-space, this requires combining powers with different signal-to-noise ratios since each has a different bandpass.

While there are an infinite number of orthogonal bases that span the survey volume, Vogeley & Szalay (1996) present a space that optimizes the signal-to-noise for δ^2. Their *Karhunen-Loève* (KL) basis offers better accuracy and resolution for $P(k)$ than conventional methods and more elegantly accounts for survey masks. It also produces measurements of power with uncorrelated error bars and maximum information retention.

Using the Karhunen-Loève approach requires partitioning the survey volume into a countable number of cells, each of which serves as a single dimension. The downside of calculating power this way is computational cost. In evaluating the power spectrum of the LRGs, Tegmark et al. (2006) describe the process as "numerically painful to implement and execute" and requiring "about a year of CPU time."

Our power spectra in this dissertation function mostly as diagnostics of other processes, and not as conclusive measures of galaxy clustering. Consequently, we do not exercise the full rigor of the Tegmark team, especially as it relates to setting error bars. We do, however, acknowledge the merits of the KL framework. The chapters that follow will utilize discretized spaces and rotated coordinate systems, partly with an eye towards extending this analysis at a later point in time. (More on the mechanics of the Karhunen-Loève transform

will be provided in §4.3.)

Discretization might seem at odds with the goal of calculating the clustering power spectrum at arbitrarily high resolution, but there are already limiting factors. First, unlike in an infinite Universe, a finite-sized survey will not have galaxies separated by every distance $r \geq r_0$ where r_0 is the physically imposed minimum intergalactic separation. Consequently, not every scale k will be represented in the galaxy power spectrum. The division of space into either Fourier gridboxes for use in a Fast Fourier Transform (FFT), or cells as with the KL basis imposes further resolution limits on the scale of the size of those regions.

The conventional response to this limitation has been to calculate $P(k)$ by averaging k-modes over bins where $k - \Delta k/2 < k < k + \Delta k/2$. The minimum bin spacing is set to be $\Delta k = 2(2\pi/2L)$ where the Fourier box runs from $[L, L]$ in all three dimensions. Because proximate length scales are inevitably averaged over anyway, grouping nearby galaxies into discretized volumes and comparing the enclosed counts between those volumes is arguably preferable to counting pairs one at time.

We remind the reader that our primary goal is to introduce improved statistical techniques to handle low level noise. The data cleansing method we describe in this dissertation takes a Bayesian perspective on this problem and requires the employ of a prior probability distribution for the power spectrum. A best-guess function of this kind is referred to as a *fiducial power spectrum*, $P_{fid}(k)$.

Because our overdensity signal will ultimately be simulated over discrete, spherical volumes (see §4.1), discontinuities between adjacent regions of space wipe out high frequency

components. In signal processing this is conventionally handled by convolving the over-density field with a window function $W(r)$ appropriate to the sampling geometry of the survey. (Equivalently, $\xi(r)$ would be convolved with the square of the window function.) Convolutions in real-space are products in Fourier space such that the power spectrum we use in equation (3.19) will actually be

$$P(k) = |W(k)|^2 P_{fid}(k). \tag{3.22}$$

The precise form of $|W(k)|^2$ used in our simulations is derived in §4.4.

The fiducial power spectrum that initiates our analysis is a *mass* density power spectrum derived from the results of the Wilkinson Microwave Anisotropy Probe (WMAP). Simulating MGS galaxy clustering requires a *galaxy* density field. Intuitively one might imagine that since galaxies are comprised of mass the clustering properties of the two would the same. While they are similar, the exact truth is slightly more complicated.

3.2.3 Bias

While a mass density field is continuous, a galaxy density field is discrete. A volume of space either contains an MGS galaxy, or it does not. There exists some mass threshold above which all MGS targets reside and below which none reside. MGS galaxies with masses above the threshold are each counted as a single object, even if one is many times more massive than another. All other volumes, including those with large masses corre-

sponding to non-MGS objects, are not counted at all. From a signal processing perspective, galaxies are "clipped" tracers of mass. MGS mass density peaks above the threshold behave like step functions while everything else is zeroed out.

A problem of this sort was worked out by Bell Labs scientists in the 1950's. They were attempting to remove telephone noise with variance σ^2 through a clipping procedure. They set their threshold amplitude to a multiple of the variance such that $A_{threshold} = \nu\sigma$ and discovered that if the underlying signal had a correlation function $\xi(r)$, then the clipped correlation function was $\xi_{th}(r) \approx \nu^2\xi(r)$.

When dealing with galaxy density fields, cosmologists use a *bias factor* b instead of ν (see e.g. Bardeen et al., 1986). The implication is that the galaxy power spectrum $P_g(k)$ is amplified by a factor of b^2 from the mass power spectrum $P_m(k)$,

$$\delta_g \sim b\,\delta_m \quad \Leftrightarrow \quad \xi_g(r) \sim b^2\xi_m(r) \quad \Leftrightarrow \quad P_g(k) \sim b^2 P_m(k). \tag{3.23}$$

Different classifications of galaxies have different thresholds, so the bias factor depends on the characteristics of one's galaxy sample. From the SDSS we have learned that older, redder elliptical galaxies have $b \sim 1.9$, while younger, bluer spiral galaxies (which have a lower noise variance) have $b \sim 1.2$. The MGS falls into the latter category so we will adopt a bias factor of $b = 1.2$ when converting between mass and galaxy density fields. More complicated biasing models include a scale dependency within b, but this level of detail is not required for our purposes and is ignored in this analysis.

3.2.4 Motivation

Substantial efforts have been made to measure cosmological parameters (e.g. Tegmark et al., 2004) or parameterize matter/galaxy simulations (e.g. "Coyote Universe" simulation suite of Lawrence et al. (2010)) by matching results to the observed power spectrum. Photometric zero-points, as we shall see in §7.5.2, introduce power perturbations on all length scales. Given the variety of measurements that depend upon a high-precision power spectrum, effectively handling small errors such as these is key.

The amount of information that can be gained from a precise measurement of the power spectrum is abundant. The Universe's matter density depends upon the characteristics of the cosmological horizon (i.e. the largest distance from which information can be received) at a time when it was equal parts matter and radiation, otherwise known as *matter-radiation equality*. The signature of this horizon is embedded within the large-scale power spectrum (Percival et al., 2007). Bias present on large scales provides information about primordial non-Gaussianity (Dalal et al., 2008). The shape of $P(k)$ also depends upon a "scalar index" n_s (Chung et al., 2003) that describes how density fluctuations vary with scale. The spectrum's amplitude is quantified using σ_8, the magnitude of fluctuations on scales of $8\,h^{-1}\mathrm{Mpc}$.

The primordial interaction between photon pressure and sound waves left imprints in the power spectrum in the form of a series of shallow peaks and valleys (e.g. Blake et al., 2011). These oscillations occur at characteristic length scales related to the physical densities of cold dark matter $\Omega_c h^2$ and baryons $\Omega_b h^2$, making the power spectrum a valuable

tool for determining mass densities and the Hubble parameter (e.g. Peebles & Yu, 1970; Sunyaev & Zeldovich, 1970; Doroshkevich et al., 1978).

These cosmological parameters can also be quantified through the cosmic microwave background (CMB) radiation. Yet in order to break CMB degeneracies, it is crucial that there exist other ways to test cosmological models. The galaxy power spectrum offers such an alternative, but its parameter constraints are generally less precise than the CMB's. Much can be achieved, then, by finding new ways to reduce statistical and systematic uncertainties in $P(k)$.

In combination with CMB measurements of the sound horizon at the baryon drag epoch, the BAOs also provide a standard cosmological ruler that can place constraints on dark energy (e.g. Blake & Glazebrook, 2003; Hu & Haiman, 2003; Seo & Eisenstein, 2003). The ruler may be applied both radially and tangentially to measure the redshift-dependent Hubble parameter and angular diameter distance. On scales of about $100\,h^{-1}\mathrm{Mpc}$, BAOs can also measure the curvature of the Universe, telling us about its expansion history (Blake et al., 2007).

A large-scale hemispherical asymmetry in the CMB has been observed by both Planck (Akrami et al., 2014) and WMAP (e.g. Eriksen et al., 2007). This lopsidedness has been dubbed "the axis of evil" and hints at a potential weakness of the prevailing inflationary paradigm. Improved, high-precision measurements of the power spectrum can help constrain possible explanations.

As future chapters will demonstrate, our primary path towards precision cosmology in-

volves optimally measuring overdensities in cells. Given all this, it is fair to wonder why we don't attempt to estimate the power spectrum (or cosmological parameters) directly. In addition to the inherent benefit of having multiple lines of analysis with which to corroborate our conclusions, overdensities are used to calculate more than the just the linear power spectrum $P_\delta(k)$.

Unwanted covariances between Fourier modes can be reduced using the power spectra of the log densities $P_{\ln(1+\delta)}(k)$ and Gaussianized densities $P_{\mathrm{Gauss}(\delta)}(k)$, both of which derive from δ. These spectra can also be more effective in extracting information from the matter field, affecting measurements of parameters like tilt (Verde & Peiris, 2008) and neutrino mass (Zhao et al., 2013; Swanson et al., 2010). They can also greatly reduce the nonlinearities in the dark matter power spectrum and capture more information than the linear spectrum on translinear scales (Yu et al., 2011; Neyrinck et al., 2009, 2006; Rimes & Hamilton, 2005). Gaussianized spectra have also been used to accurately reconstruct maps of initial fluctuations from fully sampled, sparsely sampled, and biased data (Weinberg, 1992).

3.3 The Relationship Between the Correlation Function and the Power Spectrum

We turn our attention to the relationship between the correlation function $\xi(r)$ and its associated power spectrum $P(k)$,

$$\begin{aligned}\xi(r) &= \langle \delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2)\rangle \\ &= \langle \frac{1}{(2\pi)^3}\int d^3k\,\delta(\boldsymbol{k})\,e^{i\boldsymbol{k}\cdot\boldsymbol{x}_1}\frac{1}{(2\pi)^3}\int d^3k''\,\delta(\boldsymbol{k}'')e^{i\boldsymbol{k}''\cdot\boldsymbol{x}_2}\rangle. \end{aligned} \tag{3.24}$$

Transforming variables such that $-\boldsymbol{k}' = \boldsymbol{k}''$,

$$\xi(r) = \langle \frac{1}{(2\pi)^6}\int d^3k\,\delta(\boldsymbol{k})\,e^{i\boldsymbol{k}\cdot\boldsymbol{x}_1}\int -d^3k'\,\delta(-\boldsymbol{k}')e^{-i\boldsymbol{k}'\cdot\boldsymbol{x}_2}\rangle. \tag{3.25}$$

Because $\delta(\boldsymbol{x})$ is real, $-\delta(-\boldsymbol{k}') = \delta^*(\boldsymbol{k}')$ via equation (3.13) and

$$\begin{aligned}\xi(r) &= \frac{1}{(2\pi)^6}\iint d^3k\,d^3k'\,e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_1 - i\boldsymbol{k}'\cdot\boldsymbol{x}_2}\langle\delta(\boldsymbol{k})\,\delta^*(\boldsymbol{k}')\rangle \\ &= \frac{1}{(2\pi)^6}\iint d^3k\,d^3k'\,e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_1 - i\boldsymbol{k}'\cdot\boldsymbol{x}_2}(2\pi)^3P(k)\delta^{(3)}(\boldsymbol{k}-\boldsymbol{k}') \\ &= \frac{1}{(2\pi)^3}\int d^3k\,e^{i\,\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}P(k). \end{aligned} \tag{3.26}$$

This integral is most conveniently evaluated in spherical coordinates. There is no depen-

dence on the azimuthal angle, so a factor of 2π can be pulled out. It follows,

$$\begin{aligned}\xi(r) &= \frac{1}{4\pi^2}\iint dk\, d\varphi\, k^2 e^{ikr\cos\varphi} P(k)\sin\varphi \\ &= \frac{1}{2\pi^2}\int dk\, k^2 \frac{\sin(kr)}{kr} P(k) \\ &= \frac{1}{2\pi^2}\int dk\, k^2 j_0(kr) P(k), \end{aligned} \tag{3.27}$$

where j_0 is the first spherical Bessel function. This result is not specific to cosmology, but is a more general result of Fourier theory for an isotropic function. By working through the process in reverse, it can also be shown that

$$P(k) = 4\pi \int dr\, r^2 \frac{\sin(kr)}{kr}\xi(r). \tag{3.28}$$

Because $P(k)$ depends quadratically on δ, it contains all the information needed to specify δ's second moment statistics. If the probability distribution of $\delta(r)$ follows a multivariate Gaussian form, as inflationary theory predicts and observations appear to suggest, then $P(k)$ is even more powerful — it can answer *any* statistical question about δ. Suffice it say, the power spectrum is one of cosmology's most important functions.

In an infinite Universe, knowing the two-point correlation function $\xi(r)$ with complete knowledge means knowing the power spectrum $P(k)$ identically and completely, and vice versa. In the real world there are practical subtleties like finite volumes and edge effects that affect how the errors propagate. The best we can do is to evaluate estimates of $\hat{\xi}$ and

$\hat{P}$, which are not exact Fourier transform pairs. This means that given a particular survey it might be advantageous to measure a power spectrum, while in other cases a correlation function may be preferable.

Deciding which statistic is preferable for one's purposes involves knowing what each measures. The two-point correlation function measures the excess probability of finding two galaxies separated by a distance r, while the power spectrum measures the relative density contributions on different scales. The latter is a quantity that falls more readily out of inflationary theory, and so in a sense can be considered more "natural."

The correlation function is more frequently used on small scales where high resolution is required. Also, as we will show in a future section, it is more sensitive to uncertainties in the mean density, scaling as n^{-2} while the power spectrum scales with n^{-1}. The true power spectrum $P(k)$ is strictly positive which may help in interpretations of $\hat{P}(k)$. This same requirement in the correlation function takes the form of complicated integral constraints.

3.4 The Redshift-Space Correlation Function

In observational astronomy the term *real-space* (or *configuration-space*) refers to a mapping of galaxy positions as they actually are. The term *redshift-space* (or *observer-space*) refers to the mapping of those same objects at positions derived from their redshifts. When galaxies' recession velocities depend only on the Hubble flow, these two spaces are the same.

Frequently however, galaxies possess peculiar velocities in addition to the Hubble flow that obscure their true depths. (For a review see Hamilton (1998); Kaiser (1987); for an application to the DR7 MGS, Howlett et al. (2015).) Clustering then becomes a function of the real-space correlation function and the galaxies' pairwise velocity dispersion (Fisher, 1995; Scoccimarro, 2004).

Peculiar velocities transform real-space galaxy distributions into redshift-space. Equivalently, the underlying isotropic correlation function $\xi(r)$ transforms into a non-isotropic *redshift-space correlation function*, $\xi^{(s)}(r, \theta, \gamma)$ (Davis & Peebles, 1983). In this section, we reveal the connection between the two. In §4.4, we exploit this connection to simulate clustering signal one would actually measure in observer-space.

A real-space overdensity $\delta^{(r)}$ in a linear-regime galaxy cluster where perturbations are assumed small will induce radial peculiar velocities v_r along the line-of-sight leading to the so-called "finger-of-God" effect (Jackson, 1972). The Fourier space form of the continuity equation $\dot{\delta}_k + (ik_\alpha v_\alpha(\boldsymbol{k}))/a = 0$ relates the two for each component α once it is evolved in time through the growth factor,

$$v_r(\boldsymbol{k}) = \frac{ik_r}{k^2} H_0\, a_0\, \beta\, \delta^{(r)}(\boldsymbol{k}), \tag{3.29}$$

where

$$\beta = \frac{f}{b} \cong \frac{\Omega_m^\gamma}{b}. \tag{3.30}$$

In equations (D.9) and (D.10) $k_r = |\boldsymbol{k}|\hat{k} \cdot \hat{r} = k\mu$ is the radial component of the wave vector in Fourier space, f is a scaling related to the growth of structure in the Universe, and γ is a gravitational growth index. For a wide range of cosmological parameters $f = \Omega_m^\gamma$. Here, we set $\gamma = 0.6$, which is consistent with ΛCDM models (e.g. Peebles (1980); Linder (2005), though $\Omega_m^{4/7}$ is sometimes used as well; see Appendix A).

The net effect is that the spherically symmetric power spectrum in real-space $P(k)$ is modulated by the factor β and becomes anisotropic via the directional cosine between the wave vector and the line-of-sight μ (see e.g. Kaiser, 1987; Hamilton, 1992). After some approximations and simplifying assumptions, we get

$$P^{(s)}(\boldsymbol{k}) = P(k)\big(1 + \beta\mu^2\big)^2 . \tag{3.31}$$

When $b = 1.2$ and $\Omega_m = 0.3$, we find $\beta = 0.405$. Along the line-of-sight, $\mu = 1$ and $(1 + \beta\mu^2)^2 \approx 2$. If $\boldsymbol{k}$ is perpendicular to the line-of-sight, then $(1 + \beta\mu^2)^2 = 1$. This suggests that the MGS power spectrum in redshift-space will be deformed like a football relative to real-space. The power approximately doubles or "elongates" along the line-of-sight, while the perpendicular direction remains unaffected.

There are several ways to handle these distortions. A class of "perturbation theory" solutions focuses on density fluctuations and velocity fields in an Eulerian cosmological fluid. These include Lagrangian perturbation theory (Buchert, 1992; Bouchet et al., 1995; Bernardeau et al., 2002), integrated perturbation theory (Matsubara, 2008), and convolved Lagrangian perturbation theory (Carlson et al., 2013).

Szalay et al. (1998) (hereafter SML98) relax the simplifications of equation (D.17), introduce a more convenient geometry and present a closed-form analytic solution for the redshift space correlation function. SML98 adopt the perspective that the correlation function between two points in space is a function of the triangle those points form with the observer, as shown in Figure 3.1.

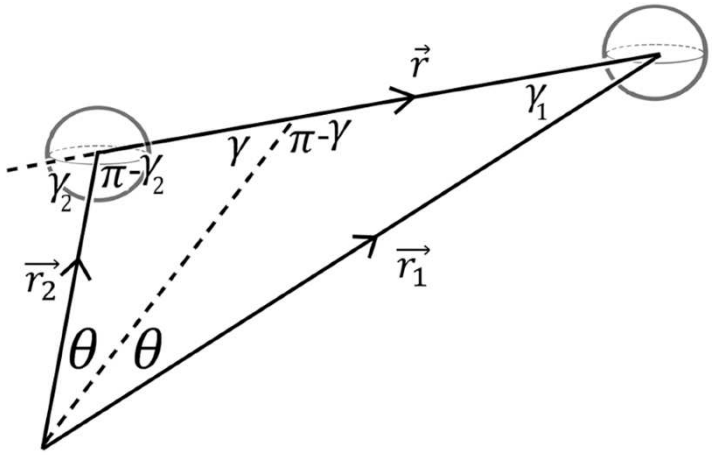

Figure 3.1: Correlation function geometry of SML98. Spheres represent two points in space separated by an angle 2θ and distance r.

This geometry facilitates expressing the correlation function in terms of $r = |\boldsymbol{r}|$. The convention is to require $r_1 \geq r_2$ and therefore $\gamma_2 \geq \gamma_1$. It can be shown that $\gamma_1 + \gamma_2 = 2\gamma$ and $\gamma_2 - \gamma_1 = 2\theta$, meaning that the entire geometry can be summarized through r, γ, and θ. The Euclidean formulas for each are straightforward. Equation (3.33) follows from the law of cosines, and equation (3.34) from the law of sines.

$$2\theta = \cos^{-1}\left(\frac{\boldsymbol{r}_1 \cdot \boldsymbol{r}_2}{r_1 r_2}\right), \tag{3.32}$$

$$r^2 = r_1^2 + r_2^2 - 2r_1 r_2 \cos(2\theta), \tag{3.33}$$

$$\gamma = \sin^{-1}\left(\frac{r_1 + r_2}{r}\sin\theta\right). \tag{3.34}$$

SML98 present the redshift space correlation function as $\xi^{(s)}(r,\theta,\gamma) = c_{00}\xi_0^{(0)} + c_{02}\xi_2^{(0)} + c_{04}\xi_4^{(0)} + c_{11}\xi_1^{(1)} + c_{13}\xi_3^{(1)} + c_{20}\xi_0^{(2)} + c_{22}\xi_2^{(2)}$, where

$$\xi_L^{(n)}(r) = \frac{1}{2\pi^2}\int dk\, k^2 k^{-n} j_L(kr) P(k). \tag{3.35}$$

Here, j_L is the spherical Bessel function and $P(k)$ is the spherically symmetric power spectrum. The coefficients are

$$c_{00} = 1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2 - \frac{4}{15}\beta^2\cos^2\theta\sin^2\theta, \tag{3.36}$$

$$c_{02} = -\left(\frac{4}{3}\beta + \frac{4}{7}\beta^2\right)\cos 2\theta\, P_2(\cos\gamma) - \frac{2}{3}\left(\beta - \frac{1}{7}\beta^2 + \frac{4}{7}\beta^2\sin^2\theta\right)\sin^2\theta, \tag{3.37}$$

$$c_{04} = \frac{8}{35}\beta^2 P_4(\cos\gamma) - \frac{4}{21}\beta^2\sin^2\theta\, P_2(\cos\gamma) - \frac{1}{5}\beta^2\left(\frac{4}{21} - \frac{3}{7}\sin^2\theta\right)\sin^2\theta. \tag{3.38}$$

In equations (3.37) and (3.38), P_l is the Legendre polynomial.

During the derivation of this correlation function, a term $\alpha(r)v_r(\boldsymbol{r})$ emerges where $\alpha(r) = [2 + \partial \ln S(r)/\partial \ln(r)]/r$ and $S(r)$ is the selection function that varies slowly in r. The velocity scale is often much smaller than the depth of the SDSS MGS survey, so this term is ignored here. All of the higher-order parameters (i.e. c_{11}, c_{13}, c_{20}, c_{22}) depend on α. Therefore, the lower-order parameters capture most of the relevant physics, and we evaluate henceforth that

$$\xi^{(s)}(r, \theta, \gamma) = c_{00}\xi_0^{(0)} + c_{02}\xi_2^{(0)} + c_{04}\xi_4^{(0)} \tag{3.39}$$

The equations above define the distance r between galaxies in the simple Euclidean sense. However, in terms of the correlation function, using a Euclidean r is not quite right. What should be measured instead is the distance between galaxies at the time corresponding to our current observations of their redshifts and angular separation, not their separation distance today. This problem was tackled by Liske (2000) who derived an improved distance measure between galaxies for use with the correlation function. The details of this measurement scheme are provided in Appendix A.

The incongruity between Liske and SML98 does present a problem. A full, derived correction to make these consistent would be extremely difficult and probably not worthwhile. Our compromise is to calculate γ using r in the simple Euclidean sense, but to use Liske's r in evaluating $\xi_L^{(n)}(r)$. Because r appears in all of the Euclidean correlation functions and γ only in the higher-order coefficients c_{02} and c_{04}, $\xi^{(s)}(r, \theta, \gamma)$ is less sensitive to errors in

γ.

We also wish to make mention of the Alcock-Paczynski (AP) effect, which reveals that the choice of one's cosmological model can induce further anisotropies into the correlation function (Alcock & Paczynski, 1979) and consequently introduce systematic bias (Ballinger et al., 1996; Simpson & Peacock, 2010). While accounting for the AP effect is certainly important in performing precision cosmology, we choose to ignore it in this analysis for a couple reasons. First, for galaxies in the redshift range of the MGS, the error it induces in f should only be on the order of 1%. Second, the efficacy of our noise cleansing method is evaluated using signal/data realizations and cleansing matrices drawn from the same, fixed cosmological model.

We conclude by mentioning that there are other methods by which a two-point correlation function can be generated. This section described the process of doing so using a fiducial power spectrum impacted by z-space effects as reflected through a cosmological model parameterization. In §6.2, we will utilize an empirical 2PCF estimator to create a model correlation function that better matches a subsample of the MGS.

Chapter 4

Signal and Noise in a Discretized Space

This chapter will explain how to generate realizations of clustering signal, shot noise, and systematic zero-point noise in a discretized space. Our discretization strategy will be to divide the SDSS survey volume into a collection of tens of thousands of densely packed spherical cells. We explain how to set their positions and sizes. We discuss the how the cells project onto regions like SECTORs and PRIMARY SEGMENTs, and provide references regarding the calculation of their intersection volumes.

We briefly summarize principle component analysis and show how it can be applied towards generating mock Universes without noise. We apply the Schechter function results to calculate the expected number of galaxies in each cell as a way to quantify shot noise. Using the cell/PRIMARY SEGMENT intersections and assumptions about the distribution of photometric zero-points, we illustrate how sample systematic noise vectors can be generated.

Each spherical cell i can be considered a separate dimension $\hat{\boldsymbol{e}}_i$ of *cell-space*. Yet it turns out that many interesting solutions for expected signal and noise are best calculated in rotated spaces. This dissertation utilizes a number of them, so a section that summarizes all required notation is offered for convenience. The chapter concludes with a census of the single-cell variances of each simulated signal and noise component.

4.1 Cell Geometry

We choose to divide space into a set of nonoverlapping spherical cells. While cubic cells would fill a greater volume, they greatly complicate evaluation of the correlation function. Viewed from the perspective of a single cube, the appearance of all other cubes is nonuniform. Some would appear face-on, others would have their vertexes be along the line-of-sight of their centers, and the majority would be something in between. This would introduce yet another break from isotropy above and beyond what the redshift-space correlation function already demands.

We need not make our lives that needlessly complicated. The more elegant option is using tightly packed spheres. Spheres possess isotropic geometry when viewed from any angle, which resolves the difficulty with the correlation function. Spheres do have the downside of not filling all space, which means some galaxies will go uncounted. When handling hundreds of thousands of objects over length scales that will be inevitably averaged, this is actually not so great a concern. Should one think otherwise, the spheres

can be permitted to overlap provided appropriate steps are taken to account for the added cross-correlations.

The cells should occupy as much of the survey's volume as one wishes to probe. The number of (nonoverlapping) cells, therefore, becomes a function of their radius. Larger cells wash out small-scale behavior while smaller cells increase computation costs. A careful balance must be struck and should be based on one's needs and computational resources.

To pack spheres as tightly as possible, the ideal approach is the Hexagonal Closest Packing (HCP) arrangement. Nonoverlapping spheres positioned via the HCP will fill about 74% of the available volume. It operates by placing spheres at every linear combination of three unit vectors:

$$\boldsymbol{a} = (1, 0, 0)\,, \quad \boldsymbol{b} = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}, 0\right), \quad \boldsymbol{c} = \left(\frac{1}{2}, \frac{\sqrt{3}}{6}, \frac{\sqrt{6}}{3}\right). \tag{4.1}$$

These vectors are scaled by the separation distance between adjacent spheres' centers, or twice the radius r. The positions of cell centers can be written as $\boldsymbol{d} = 2r(n_x\boldsymbol{a} + n_y\boldsymbol{b} + n_z\boldsymbol{c})$, where n_x, n_y, and n_z are integers.

We generate three sets of cells with radii: 7, 11, and 16 $h^{-1}\mathrm{Mpc}$ which we refer to as the R7, R11, and R16 cases. The former two groups of cells are placed at redshifts $0.02 \leq z \leq 0.22$ where the upper limit is chosen because a) the selection function drops to less than 1% for $z > 0.22$ and b) we needed to limit the total number of cells in the survey. The $16\,h^{-1}\mathrm{Mpc}$ radius cells, which will be less plentiful in number, are extended to $z < 0.3$

Radius (h^{-1}Mpc)	Number	Redshift Range	Comoving Distance Range (h^{-1}Mpc)
7	78,845	$0.02 < z < 0.22$	$60 < d_C < 626$
11	19,737	$0.02 < z < 0.22$	$60 < d_C < 626$
16	15,166	$0.02 < z < 0.30$	$60 < d_C < 836$

Table 4.1: Spatial properties of discretized cells in comoving space.

to more effectively probe the effect of noise at high redshifts. Table 4.1 summarizes our sets of spheres.

We require that each cell have at least 62% of its volume within the spectroscopic footprint. The complicated nature of the cell/region intersections precludes measuring those volumes analytically. Instead, we adopt a numerical approach by generating uniformly distributed angular random variables (aka *angular randoms*) within the cell's footprint. Each variable's line-of-sight passes through a different length of the cell given by $l = 2\sqrt{r^2 - \chi^2(1-c^2)}$, where l is the length of the chord, r is the radius of the cell, χ is the comoving distance to the cell's center, and $c \equiv \cos\theta = \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{x}}$ is the cosine of the angle between the vector pointing to the center of the sphere, $\hat{\boldsymbol{n}}$, and the direction directed towards the point at random ray entered, $\hat{\boldsymbol{x}}$. Figure 4.1 shows the ratio $l/(2r)$ for a randomly selected cell.

The set of angular randoms is used to calculate β_{spec}, the cell's fractional volume within the spectroscopic footprint. If W_i represents the cumulative lengths of all chords passing through the i^{th} cell, and w_i represents the cumulative lengths of those that *also* pass within the spectroscopic footprint, then $\beta_{spec} = w_i/W_i$ where $0 < \beta_{spec} \leq 1$.

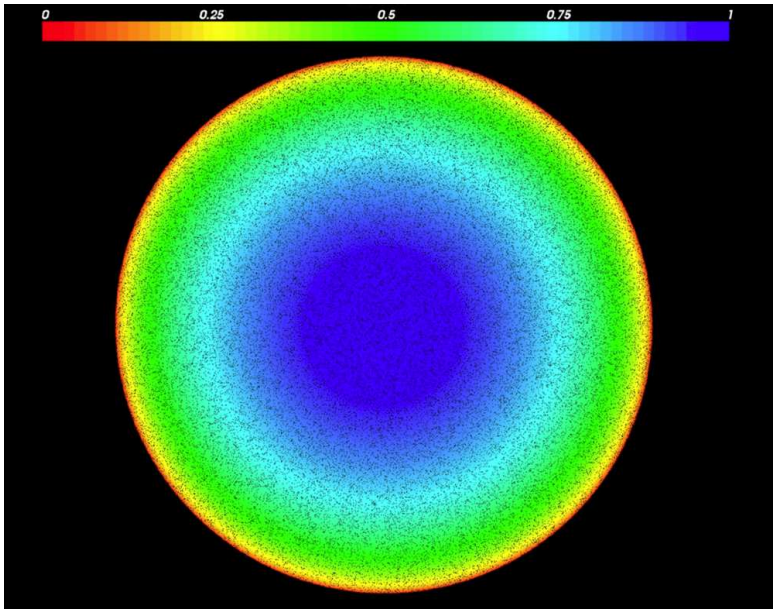

Figure 4.1: Visual representation of chord weights through a randomly selected sphere. Weights are denoted by the color bar. Angular random variables that pass directly through the center of the sphere are assigned a weight of one. Those that are tangential to the sphere receive weights of zero.

Finding the volume each PRIMARY SEGMENT intersects is accomplished in a similar way. The footprint of the PRIMARY SEGMENT is superimposed upon the projection of the cell so that each angular random variable can be assigned to a particular PRIMARY SEGMENT, or otherwise will pass outside the footprint. Enough randoms were passed through the cells to ensure the volume fractions were known to better than 1% accuracy. The relationship between photometric and spectroscopic footprint volumes is plotted in Figure 4.2.

The process of generating high-density angular randoms in the appropriate regions to calculate volume fractions of cell/region intersections was nontrivial. The computational cost of searching over SDSS regions for all cells was expensive and care had to be taken

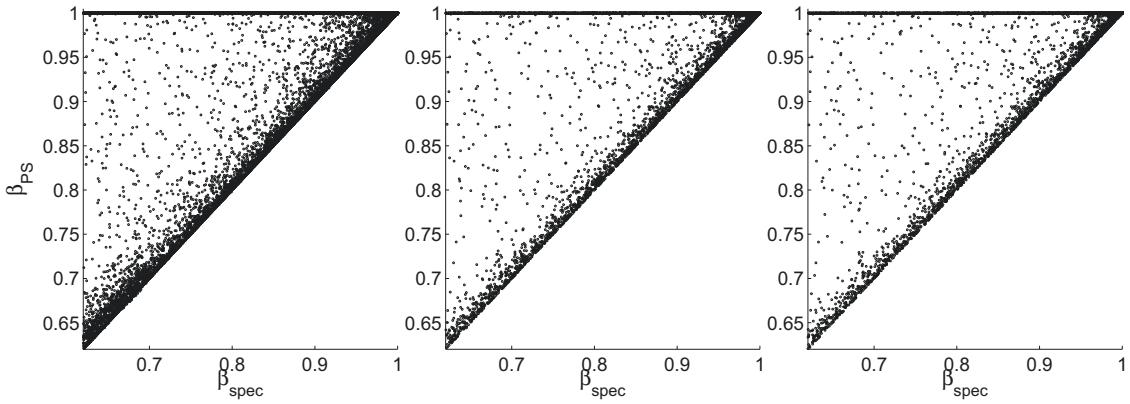

Figure 4.2: Comparison between the volumes of cells inside the improved photometric footprint, β_{PS}, and improved spectroscopic footprint, β_{spec}. Cases presented are R7 (*left*), R11 (*middle*) and R16 (*right*). Cells along the horizontal line for which $\beta_{PS} = 1$ lie entirely within the photometric footprint but can reach outside the spectroscopic footprint. Cells along the unit slope lie in regions where the photometric and spectroscopic footprints overlap.

to make the calculations as efficient as possible. For more details on the theory of angular randoms, the weighting function, region filtering, method of calculating region volumes, and the 62% criterion, please reference Appendix E.

A set of discrete cells generated and characterized in this fashion acts as a standard basis where the unit vector $\hat{\boldsymbol{e}}_i$ equals 1 in the i^{th} cell and zero otherwise. Any vector $\boldsymbol{\nu}$ expressed as a linear combination of values in cells, $\boldsymbol{\nu} = \sum_{i=1}^{N} \nu_i \hat{\boldsymbol{e}}_i$, is said to be expressed in *cell-space*. Quantities like galaxy counts and overdensities are measured in cell-space, and as such, cell-space ought to be thought of as the "default coordinate system." In future sections we will introduce other coordinate systems (e.g. signal-space, W-space) that are more convenient for specific applications.

4.2 Power Spectra

The translation from $\hat{\boldsymbol{e}}_i$ in configuration-space to $\hat{\boldsymbol{e}}_i(\boldsymbol{k})$ is accomplished through an FFTW algorithm that optimizes to the user's hardware. It requires that the survey cells be encased within a cubic *Fourier box* with sides spanning from $[-L, L]$. To promote periodicity and reduce aliasing effects (i.e. power from one frequency band getting mapped to another) the Fourier box should ideally be twice the size of the largest structures in the survey.[1] From Table 4.1 this corresponds to $L = 1252\,h^{-1}\mathrm{Mpc}$ and $L = 1672\,h^{-1}\mathrm{Mpc}$ for R7/R11 and R16 respectively.

The Fourier box is divided into n_x cubic *gridboxes* in each dimension for a total of n_x^3. To optimize speed, it is recommended that n_x be a power of 2, like 128, 256, or 512. Smaller gridboxes are preferred for their ability to probe smaller scale structures, but require more time to compute and disk space to store.

There are no details in our system smaller than the cell's diameter, $2r$. This sets the Nyquist frequency $k_{Nyquist} = 2\pi/(2r)$ and minimum sampling rate $k_s = 2k_{Nyquist}$ required to faithfully reproduce the signal. Table 4.2 summarizes the sampling frequencies required for each cell size. For the given L, at least 512 gridboxes per dimension are needed for R7 and R11, while 256 will suffice for R16.

[1] An alternative to doubling the Fourier box size is convolving the overdensity function with a Hamming window. The convolution enhances periodicity but introduces k-space smoothing on small scales.

	Cell Scale (h^{-1}Mpc)	$k_{Nyquist}$ (h^{-1}Mpc)	k_s (h^{-1}Mpc)
R7	14	0.449	> 0.898
R11	22	0.286	> 0.571
R16	32	0.196	> 0.393

Table 4.2: Minimum sampling frequencies required to interpolate values in cells back to their continuous values. The smallest resolvable scale is taken to be the distance between adjacent cells' centers, or twice the radius. Under ideal circumstances, gridboxes will have sizes smaller the minimum k_s.

4.3 Principle Component Analysis

In cell-space the correlations between random variables (e.g. signal or noise overdensities) are represented through a covariance matrix with a dimensionality N equal to the number of cells. Covariance matrices, being positive-definite, can be factored into a product of its eigenvectors and eigenvalues in a process known as *diagonalization*.

The eigenvectors of a diagonalized matrix act as a rotated basis along which possibly-correlated random variates are no longer correlated. The process of using these eigenvector/eigenvalue pairs, otherwise known as *eigenmodes*, to extract statistical information about the underlying process is referred to as *principle component analysis* (PCA). The representation of the process along the eigenmodes, or *principle components*, is also known as the *Karhunen-Loève (KL) transform*, the benefits of which were introduced in §3.2.2.

To summarize the properties of PCA, consider an N-dimensional random vector $\boldsymbol{\delta}$ projected along a dimension $\boldsymbol{\alpha}_1$ such that $\boldsymbol{\alpha}_1^T\boldsymbol{\delta} = \sum_{i=1}^{N} \alpha_{1j}\delta_j$. The goal of PCA is to identify the unit vector $\boldsymbol{\alpha}_1$ that maximizes $\mathrm{Var}\big(\boldsymbol{\alpha}_1^T\boldsymbol{\delta}\big) = \boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{\delta}$. Using a method of Lagrange multipliers, PCA reveals that the solution is

$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1$. That is, the dimension of greatest variance is an eigenvector of $\boldsymbol{\Sigma}$. We find that $\mathrm{Var}\left(\boldsymbol{\alpha}_1^T\boldsymbol{\delta}\right) = \boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^T\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 = \lambda$ is maximized only when λ assumes its largest value, λ_1. Consequently $\boldsymbol{\alpha}_1$ must be the eigenvector associated with the largest eigenvalue of $\boldsymbol{\Sigma}$.

By adding the constraint that the first and second principal components are uncorrelated the problem is solved again, this time for the orthogonal dimension with the largest amount of remaining variance. The solution is $\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \lambda_2\boldsymbol{\alpha}_2$. Repeating this process proves the general result — the p^{th} principle component of $\boldsymbol{\delta}$ equals the p^{th} ordered eigenvector of its covariance matrix. Furthermore, the variance along the p^{th} principle component equals the p^{th} largest eigenvalue.

In the pages that follow, this result will be used to generate random correlated signal and noise vectors. It will also reveal the underlying structure of those processes in an optimal manner. Finally, it will allow a new Bayesian noise removal process introduced in Chapter 7 to be solved analytically and applied in a manner that minimizes computation time.

4.4 Clustering Signal

The fiducial power spectrum $P_{fid}(k)$ must be filtered through a window function to account for the smoothing effect of the cells. Cells smear galaxy counts identically throughout their volumes. This justifies the use of the following spherical window function $W_R(\boldsymbol{r})$ that takes a value equal to the inverse volume of cells of radius R,

$$W_R(\boldsymbol{r}) = W_R(r) = \begin{cases} 3/(4\pi R^3) & \text{if } r \leq R \\ \\ 0 & \text{if } r > R \end{cases}. \tag{4.2}$$

A filtering in configuration-space amounts to a product in Fourier space. When j_1 is taken to be a spherical Bessel function of the first kind,

$$W_R(\boldsymbol{k}) = \int W_R(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\, d^3r = 3\frac{j_1(kR)}{kR}. \tag{4.3}$$

The smoothed power spectrum $P(k) = |W_R(k)|^2 P_{fid}(k)$ is used in equation (3.35) to calculate the redshift-space correlation functions. In principle, $\lim_{R\to\infty} \int_0^R r^2 \xi_L^{(n)}(r)\, dr$ should converge, but in practice, rounding errors introduced by the Bessel functions induce oscillations at $r > 280\, h^{-1}\mathrm{Mpc}$ for $\xi_2^{(0)}$ and at $r > 300\, h^{-1}\mathrm{Mpc}$ for $\xi_4^{(0)}$. These are small effects, around the 1% level, but we account for them by replacing modest oscillations with a best fit 5^{th} degree polynomial. Once the oscillations become chaotic, the correlation functions are set to zero. The 0^{th} order correlation function runs numerically negative before oscillations begin, so we set $\xi_0^{(0)}(r > r') = 0$ where $\xi_0^{(0)}(r') = 10^{-5}$. The three relevant correlation functions are plotted in Figure 4.3. Larger windows (i.e. higher R) smear more structure, driving $\xi_L^{(0)}(r)$ downward as a general rule.

We represent clustering signal in cell-space with the N-dimensional column vector $\boldsymbol{\kappa}$. The redshift-space galaxy clustering covariance matrix, or simply the *signal matrix*, is defined as $\boldsymbol{\Sigma}_\kappa \equiv \mathrm{cov}\big(\boldsymbol{\kappa}^T\boldsymbol{\kappa}\big)$. Each element in the i^{th} row and j^{th} column $\boldsymbol{\Sigma}_\kappa[i,j]$ is set

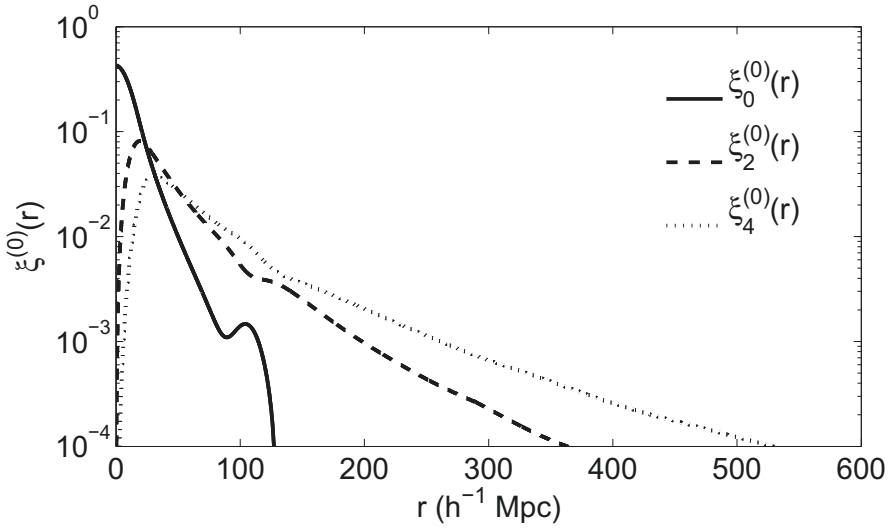

Figure 4.3: Redshift-space correlation functions from equation (3.35) convolved with the spherical window function of equation (4.2) when $R = 11\,h^{-1}\mathrm{Mpc}$. Smaller values of R drive the peak of $\xi_0^{(0)}$ upwards while larger values drive it down. Smaller values of R shift the peaks of $\xi_2^{(0)}$, and $\xi_4^{(0)}$ up and to the left while large values shift them down and to the right. The curves largely overlap otherwise.

individually by applying equation (3.39) to cells i and j which are separated by r, θ, and γ.

The diagonalization of $\mathbf{\Sigma}_\kappa$ yields

$$\mathbf{\Sigma}_\kappa = \mathbf{Z}\mathbf{\Lambda}^{(\kappa)}\mathbf{Z}^T, \tag{4.4}$$

where the i^{th} column of the orthonormal matrix $\mathbf{Z}$ contains the i^{th} eigenvector $\hat{\boldsymbol{z}}_i$ of $\mathbf{\Sigma}_\kappa$, and the i^{th} diagonal element of $\mathbf{\Lambda}^{(\kappa)}$ equals $\mathbf{\Sigma}_\kappa$'s i^{th} eigenvalue $\lambda_i^{(\kappa)}$. Eigenvectors are positioned in $\mathbf{Z}$ like so,

$$\mathbf{Z} = \begin{bmatrix} | & & | \\ \hat{\mathbf{z}}_1 & \cdots & \hat{\mathbf{z}}_N \\ | & & | \end{bmatrix}. \tag{4.5}$$

Hereafter, we will refer to the orthonormal set of basis vectors contained in $\mathbf{Z}$ as the *signal basis* which spans *signal-space*. Associated eigenvalue/eigenvector pairs will be referred to as the *signal eigenmodes*. Signal eigenmodes are indexed according to their eigenvalues from largest ($i = 1$) to smallest ($i = N$).

As an orthogonal transformation, $\mathbf{Z}$ acts as a matrix that rotates vectors in cell-space to signal-space and vice versa through $\boldsymbol{S} = \mathbf{Z}^T\boldsymbol{\kappa}$ and $\boldsymbol{\kappa} = \mathbf{Z}\boldsymbol{S}$. As an orthonormal matrix, the columns of $\mathbf{Z}$ comprise a full basis that spans cell-space. Therefore any signal vector $\boldsymbol{\kappa}$ can be represented as a linear combination of them,

$$\boldsymbol{\kappa} = \sum_{i=1}^{N} S_i \hat{\boldsymbol{z}}_i, \tag{4.6}$$

where the expansion coefficient S_i equals the projection of the signal in cell-space onto the i^{th} signal mode, $S_i = \boldsymbol{\kappa} \cdot \hat{\boldsymbol{z}}_i$.

In this Karhunen-Loève basis $\mathbf{Z}$, the signal coefficients are mean zero, $\langle S_i \rangle = 0$, and statistically orthogonal,

$$\langle S_i S_j^* \rangle = \begin{cases} \lambda_j^{(\kappa)} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{4.7}$$

This enables the generation of random signal vectors $\boldsymbol{\kappa}^{(\tau)}$.[2] We assume the signal coefficients are drawn from a Gaussian distribution, $S_i \sim \mathcal{N}(0, \lambda_i^{(\kappa)})$, such that the τ^{th} realization is

$$\boldsymbol{\kappa}^{(\tau)} = \sum_{i=1}^{N} S_i^{(\tau)} \hat{\boldsymbol{z}}_i. \tag{4.8}$$

The total signal variance equals $\sum_{i=1}^{N} \lambda_i^{(\kappa)} = \mathrm{tr}\big(\boldsymbol{\Lambda}^{(\kappa)}\big)$. We find,

$$\sum \lambda_{R7}^{(\kappa)} = 11.78 \times 10^4, \quad \sum \lambda_{R11}^{(\kappa)} = 2.95 \times 10^4, \quad \sum \lambda_{R16}^{(\kappa)} = 2.27 \times 10^4. \tag{4.9}$$

The difference among cell sizes is a function of their number. The variance per cell is the same for all radii and equals 1.4935.

The signal eigenvalues, or equivalently the variance contained within each of the signal eigenmodes, is plotted in Figure 4.4. All modes possess a non-negligible amount of variance. This indicates that there is little opportunity to express $\boldsymbol{\kappa}$ in fewer than N dimensions. In fact, as shown in Figure 4.5, for R7 it would require the first 62,045 modes, or 78.7% of N, to capture 90% of the variance. For R11 and R16 we need 86.6% and 88.6% of the modes, respectively.

The signal eigenvectors $\hat{\boldsymbol{z}}$ have a spatial representation as depicted in Figure 4.6. Lower-order eigenmodes are dimensions along which the greatest amount of signal variance exists.

[2] Other methods to generate signal realizations exist. One popular choice is drawing the real and imaginary components of $\boldsymbol{\kappa}(k)$ from a mean-zero, normal distribution with variance proportional to $\sqrt{P(k)}$.

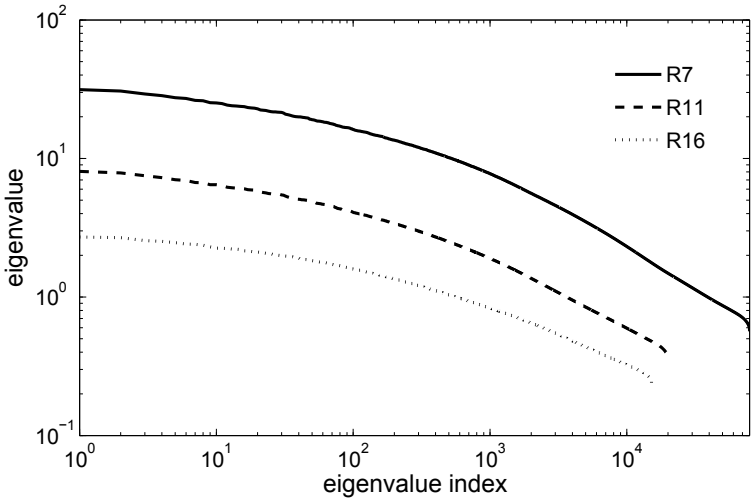

Figure 4.4: Scree plot of signal eigenvalues $\lambda^{(\kappa)}$. Eigenvalues are ranked from largest (index = 1) to smallest (index = N).

Spatially, these modes represent the largest structures possible within the SDSS footprint. Higher-order eigenmodes correspond to higher frequency, or spottier, structures.

The frequency compositions of several signal eigenmodes are presented in Figure 4.7. Each eigenmode can be characterized by the frequency at which it achieves peak power.[3] Figure 4.8 provides a census of these frequencies. While it should be possible to isolate particular signal features by targeting ranges of $\hat{\boldsymbol{z}}_i$, the same (unfortunately) cannot be said for the zero-point noise, as we will demonstrate in §4.6.

A three-dimensional visualization of some signal modes in Fourier space can be seen by clicking this link. By virtue of the Hermiticity of the Fourier transform, these modes are perfectly symmetric about the z-axis. As expected, the lowest order modes are packed in near the center. The higher order modes are roughly spherical in shape with alternating

[3]The spectrum of each mode would be closer to a Dirac delta function if not for the convolution introduced by the window function and FFT discretization.

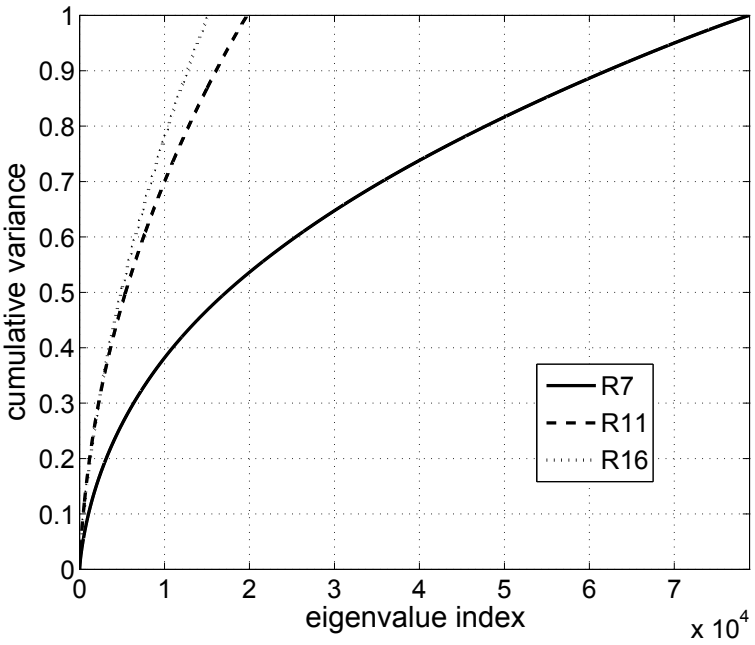

Figure 4.5: Cumulative variance of signal eigenmodes. For each eigenvalue index n, the variance reported on the vertical axis is $\sum_{i=1}^{n} \lambda_i^{(\kappa)}$.

high and low magnitude regions distributed like waves over the k-space shell.

In an attempt to reduce zero-point noise, which is mostly a large scale effect, it is tempting to isolate signal modes in the linear regime. Doing so could facilitate a dimensionality reduction and potential isolation of the noise. While $8\,h^{-1}\mathrm{Mpc}$ is typically considered the scale of nonlinearity, a value of $20\,h^{-1}\mathrm{Mpc}$ is usually used in BAO reconstruction work. In either case, Figure 4.8 reveals that none of the signal modes are ever indisputably within the linear regime, rendering a mode truncation of this kind unworkable. Fortunately, the signal structure introduced here will still play a useful role in the Bayesian noise analysis to come.

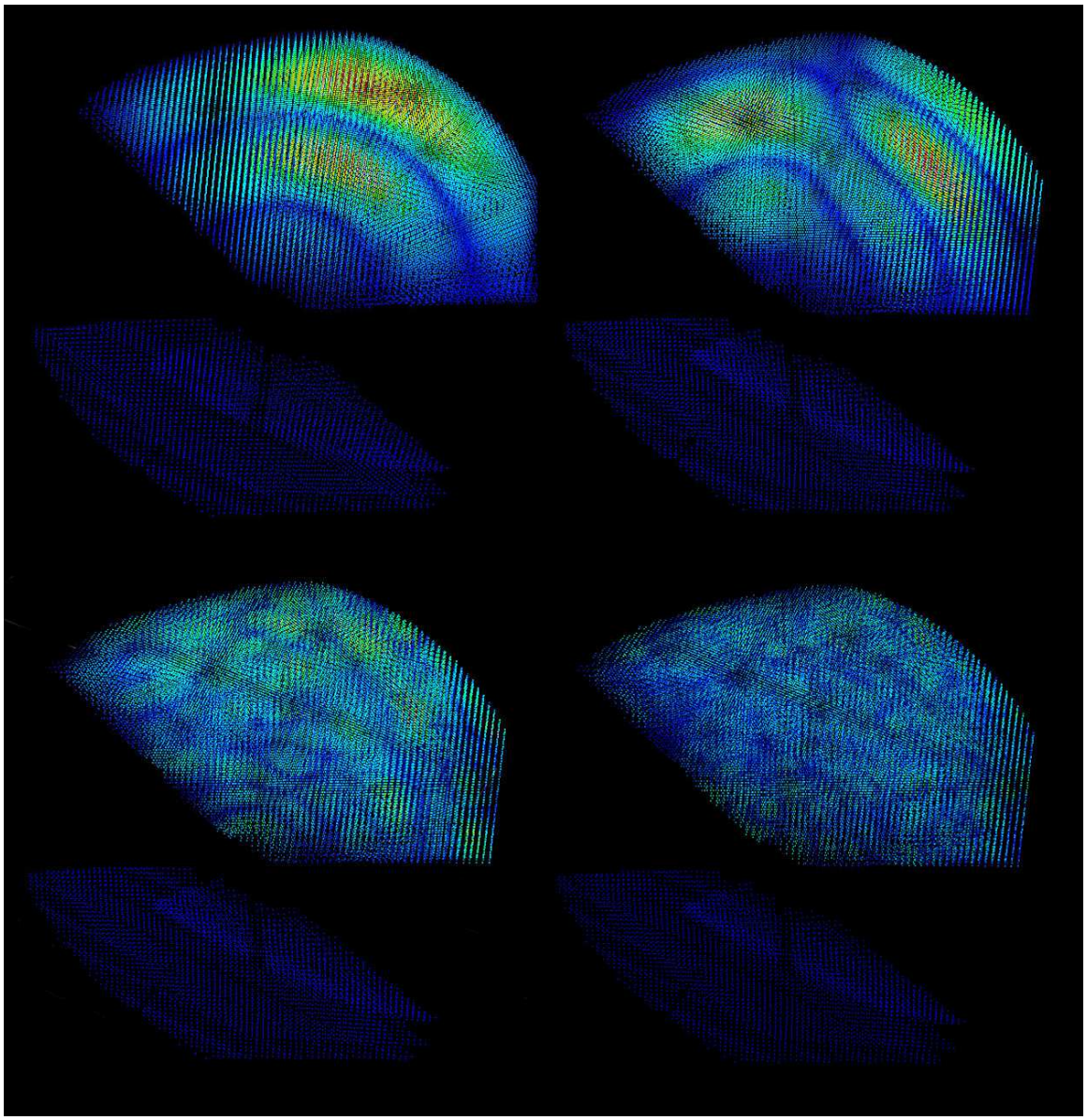

Figure 4.6: Visual depiction of four signal eigenmodes. Each pixel represents one R7 cell. Color stands for the magnitude of the eigenvector element in each cell with dark blue being of lowest magnitude and red being of highest magnitude. Starting in the upper-left corner and running clockwise these are modes 1, 4, 1000, and 200. For a slideshow containing more signal modes, visit this link.

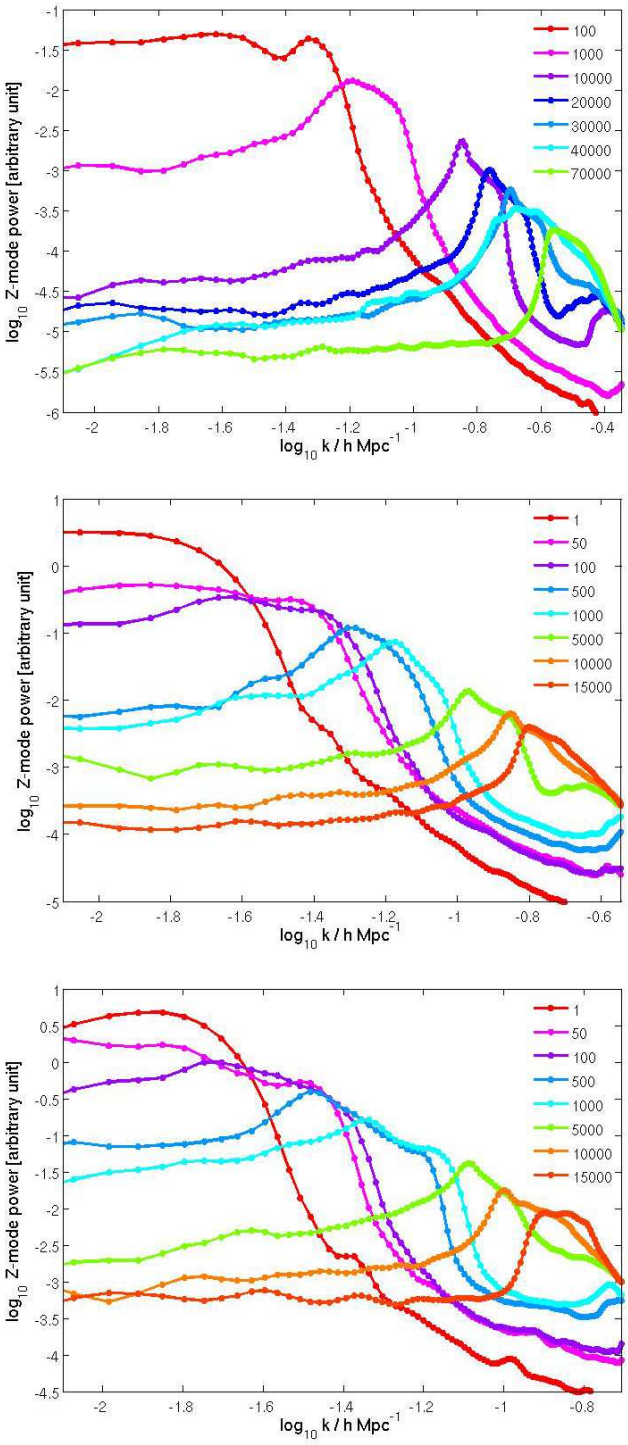

Figure 4.7: Power spectra of select signal eigenvectors $\hat{z}_i$ for R7 (*top*), R11 (*middle*) and R16 (*bottom*).

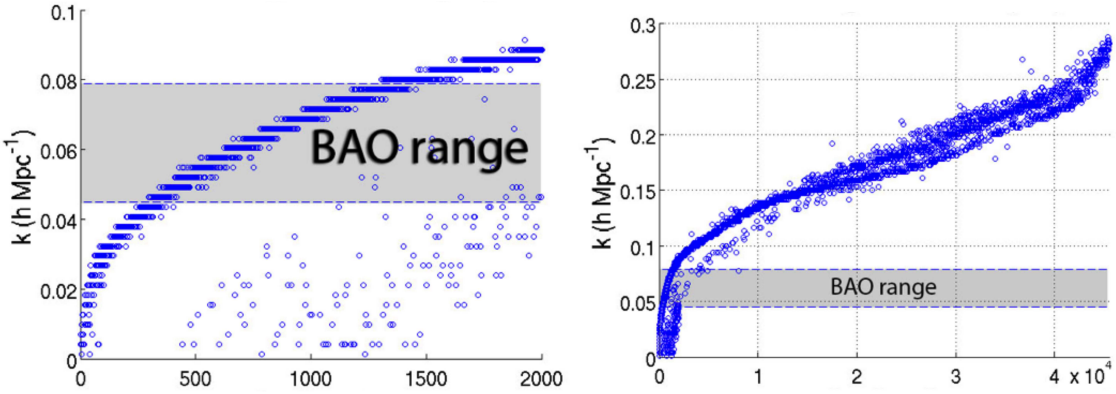

Figure 4.8: Wavenumbers principally represented by each of the first 45,000 R7 signal eigenmodes. The horizontal axis contains the indices of the eigenmodes. The values on the vertical axis are taken to be the peaks of each signal mode's power spectrum. Wavenumbers appear quantized due to the finite number of k-bins. The typical range of the BAOs, $0.045 \leq k \leq 0.079\,h\,\mathrm{Mpc}^{-1}$, is shaded in gray.

4.5 Shot Noise

We let the N-dimensional column vector $\boldsymbol{\zeta}$ represent shot noise in cell-space. The shot noise in cell i is given by ζ_i. The covariance matrix of the shot noise, or simply the *shot noise matrix*, is defined $\boldsymbol{\Sigma}_\zeta \equiv \mathrm{cov}\big(\boldsymbol{\zeta}^T\boldsymbol{\zeta}\big)$. From equation (3.10), the shot noise variance equals the reciprocal of the expected number of galaxies $\langle n_i\rangle$. We take the shot noise to be Gaussian such that $\boldsymbol{\Sigma}_\zeta$ is diagonal in cell-space and $\zeta_i \sim \mathcal{N}(0, 1/\langle n_i\rangle)$. Shot noise realizations $\boldsymbol{\zeta}^{(\tau)}$ can be drawn directly from this distribution. The diagonality of $\boldsymbol{\Sigma}_\zeta$ assumes the spheres do not overlap. If they do, adjustments to $\boldsymbol{\Sigma}_\zeta$ are needed. These are covered in Appendix F.

The number $\langle n(z_i)\rangle$ of galaxies expected within a cell center of z_i is found by integrating equation (2.16). The comoving volume out to redshift z is $(4\pi/3)\chi(z)^3$. Therefore the expected number density of galaxies at z is

$$\rho(z) = \frac{\int_z^{z+dz} n_{exp}(z')\,dz'}{(A_{spec}/A_{fs})\,\frac{4\pi}{3}\left(\chi\left(z+dz\right)^3 - \chi(z)^3\right)}. \tag{4.10}$$

The fractional area of sky taken up by the spectroscopic footprint is (A_{spec}/A_{fs}). Accordingly, n_{exp} is normalized by using all MGS targets within the spectroscopic footprint.

The i^{th} cell, centered at z_i with radius R, will reside within a redshift range $z_l \le z_i \le z_u$. Let $dV(z)$ equal the differential volume of the cell between z and $z+dz$ where $\int dV(z) = (4\pi/3)R^3$. The expected number count within cell i will then be

$$\langle n_i \rangle = \int_{z_l}^{z_u} \rho(z) dV(z). \tag{4.11}$$

The shape of $dV(z)$ is the difference between two spherical caps, the height of which is given by h, where $h = 0$ at the near end of the cell and $h = 2r$ at the far end. The volume of a spherical cap is $V = \pi h^2(3R - h)/3$. It follows that the volume of a differential cell slice is,

$$\begin{aligned} dV &= V(h + dh) - V(h) \\ &= \frac{\pi}{3}\left(3R\left((h + dh)^2 - h^2\right) + h^3 - (h + dh)^3\right). \end{aligned} \tag{4.12}$$

The function $h(z)$ is translated into a comoving distance and the volume integration occurs numerically.

Some cells extend beyond the edge of the PRIMARY SEGMENT footprint. There are no MGS targets to be counted here, pristine or otherwise. The volume fraction of cell i within the PRIMARY SEGMENT footprint is denoted $\beta_{PS,i}$ such that the expected number count therein is

$$\langle n_i \rangle = \beta_{PS,i} \langle n(z_i) \rangle. \tag{4.13}$$

The effect of shot noise is greatest where cells are small and/or the selection function is

low. The latter falls below 1% of its maximum at $z > 0.22$. The majority of R16 cells lie there, which means that shot noise has an outsize impact on this sample.

4.6 Systematic Zero-Point Noise

Systematic noise represented as ϵ_{ij} in equation (3.10) comes in many forms. Here our focus is squarely on the impact of zero-point photometric offsets on galaxy counts in cells. As discussed in §2.3.3.2, zero-point offsets Δm change the effective limiting magnitude $m_{lim} = 17.77$ of the MGS. A large enough offset can cause bright galaxies to be incorrectly excluded and dim galaxies to be incorrectly included. This changes the number of galaxies counted and, in turn, the overdensities measured in a radially-dependent way. We should expect zero-point noise to add structure that grows with distance due to the impact illustrated in Figure 2.25.

We represent zero-point noise in cell-space with the N-dimensional column vector $\boldsymbol{\eta}$. The systematic, zero-point noise covariance matrix, or simply the *zero-point noise matrix*, is defined as $\boldsymbol{\Sigma}_\eta \equiv \mathrm{cov}\big(\boldsymbol{\eta}^T\boldsymbol{\eta}\big)$. The matrix $\boldsymbol{\Sigma}_\eta$ is constructed by quantifying how offsets in PRIMARY SEGMENTs map to overdensities in cells.

Let $\langle n_{ij}\rangle$ represent the expected number of galaxies in the intersection of the i^{th} cell and j^{th} PRIMARY SEGMENT in a homogeneous Universe (i.e. absent clustering) and let n_{ij} equal the number counted if photometric offsets Δm_j are introduced. From equation (2.20) the number counted will equal $n_{ij} = \langle n_{ij}\rangle + \langle n_{ij}\rangle f_i \Delta m_j$, where $f_i \equiv f(z_i)$ and z_i is the

redshift at the center of the i^{th} cell. Summing over all N_s PRIMARY SEGMENTs j,

$$\sum_{j=1}^{N_s} n_{i,j} = \sum_{j=1}^{N_s} \langle n_{i,j} \rangle + f_i \sum_{j=1}^{N_s} \langle n_{i,j} \rangle \Delta m_j, \tag{4.14}$$

$$\frac{\sum_j n_{i,j}}{\sum_j \langle n_{i,j} \rangle} = 1 + \frac{f_i \sum_j \langle n_{i,j} \rangle \Delta m_j}{\sum_j \langle n_{i,j} \rangle}. \tag{4.15}$$

The zero-point noise in the i^{th} cell is defined as

$$\begin{aligned} \eta_i = \frac{\sum_j n_{i,j}}{\sum_j \langle n_{i,j} \rangle} - 1 &= \frac{f_i \sum_j \langle n_{i,j} \rangle \Delta m_j}{\sum_j \langle n_{i,j} \rangle} \\ &= f_i \sum_{j=1}^{N_s} p_{i,j} \Delta m_j, \end{aligned} \tag{4.16}$$

where $p_{i,j}$ equals the fractional volume the j^{th} SEGMENT occupies in the i^{th} cell,

$$p_{i,j} \equiv \frac{\langle n_{i,j} \rangle}{\sum_{j'} \langle n_{i,j'} \rangle}. \tag{4.17}$$

The method by which the fractions are calculated is covered in detailed in Appendix E

Any vector of N_s zero-point offsets $\Delta \boldsymbol{m}$ can be mapped to an N-dimensional zero-point noise vector $\boldsymbol{\eta}$ using the rotation matrix $\mathbf{A}$ where where $\boldsymbol{\eta} = \mathbf{A} \cdot \Delta \boldsymbol{m}$ and

$$\mathrm{A}_{i,j} = f_i \, p_{i,j}. \tag{4.18}$$

It follows that $\mathbf{\Sigma}_\eta = \mathbf{A}\mathbf{\Sigma}_{\Delta m}\mathbf{A}^T$, where $\mathbf{\Sigma}_{\Delta m}$ is the covariance matrix of the photometric offsets. We assume that each offset is i.i.d. with $\Delta m \sim \mathcal{N}(0, \sigma_m^2)$ such that the covariance matrix of the zero-point noise overdensities is

$$\mathbf{\Sigma}_\eta = \mathbf{A}\mathbf{\Sigma}_{\Delta m}\mathbf{A}^T = \sigma_m^2 \mathbf{A}^T. \tag{4.19}$$

As discussed in §2.1.2, the zero-point offsets result from a combination of effects across three telescopes. Through the central limit theorem, we would expect that this combination of errors is approximately Gaussian. The ubercalibrations offer the opportunity to not only test this assumption, but to quantify the true value of σ_m.

From the DR6 database, we find the difference between MGS galaxies' PT-calibrated and ubercalibrated magnitudes with the following query:

```
SELECT s.cx, s.cy, s.cz, u.petroMag_r-s.petroMag_r
FROM SpecPhotoAll s, UberCal u
WHERE s.objID = u.objID
```

When galaxy i, with a magnitude difference of Δr_{ij}, is one of n_j in PRIMARY SEGMENT j, then we calculate the offset in that region to be

$$\Delta m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \Delta r_{ij}. \tag{4.20}$$

The distribution of Δm_j's is presented in Figure 4.9. Its shape supports the approximation of Gaussianity. We find the offsets have a standard deviation of 0.0094. For simplicity, all simulations going forward will assume that $\sigma_m = 0.01$.

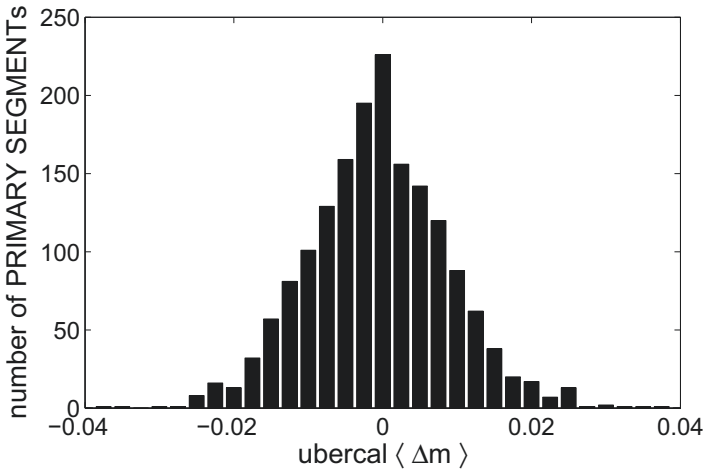

Figure 4.9: Distribution of DR6 photometric zero-points as determined through their ubercalibrations. The differences between the PT-calibrated and ubercalibrated magnitudes for MGS galaxies are averaged over their PRIMARY SEGMENTs and reported in this histogram. To avoid counting objects in very small regions, only the top 1690 PRIMARY SEGMENTs as measured by area are considered. Bins have a width of 0.0025.

Another factor that impacts zero-point overdensities is a cell's angular radius. Figure 4.10 reports how the number of PRIMARY SEGMENTs intersected by a cell decreases as a function of redshift. Because Δm_j in adjacent regions are mean zero and uncorrelated, a greater number of intersections serves to offset the percentage change in the number of counted galaxies.

In Figure 4.11, we see the standard deviation of η over all cells. A trivial result is that $\langle \sigma_\eta \rangle$ increases with σ_m. But we also see that when $\sigma_m = 0.01$, $\langle \sigma_\eta \rangle$ equals 0.0081, 0.0063, and 0.0111 for R7, R11, and R16 respectively. The zero-point errors are greater in R7 than R11 since the latter intersects more PRIMARY SEGMENTs, while both are constrained to the same redshift range. Over 56% of R16 cells exist at $z > 0.22$ where they are subjected to larger values of $f(z)$. When all cells are considered, this factor tends to dominate over

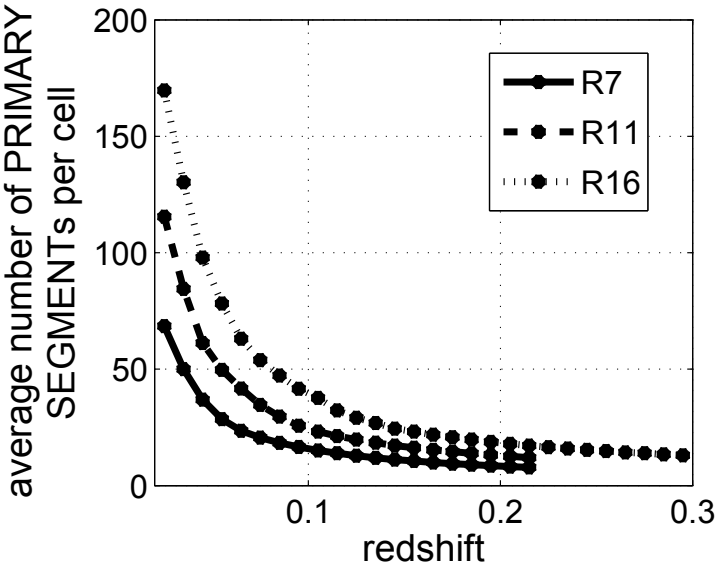

Figure 4.10: Average number of DR6 PRIMARY SEGMENTs intersecting R7, R11, and R16 cells. The numbers of intersections are averaged in redshift bins of width $\Delta z = 0.01$. Only PRIMARY SEGMENTs with physical, nonzero areas are considered.

R16's larger angular radii.

The diagonalization of $\mathbf{\Sigma}_\eta$ yields

$$\mathbf{\Sigma}_\eta = \mathbf{U}\mathbf{\Lambda}^{(\eta)}\mathbf{U}^T, \tag{4.21}$$

where the i^{th} column of the orthonormal matrix $\mathbf{U}$ contains the i^{th} eigenvector $\hat{\boldsymbol{u}}_i$ of $\mathbf{\Sigma}_\eta$ and i^{th} diagonal element of $\mathbf{\Lambda}^{(\eta)}$ equals $\mathbf{\Sigma}_\eta$'s i^{th} eigenvalue $\lambda_i^{(\eta)}$. Eigenvectors are positioned in $\mathbf{U}$ like so,

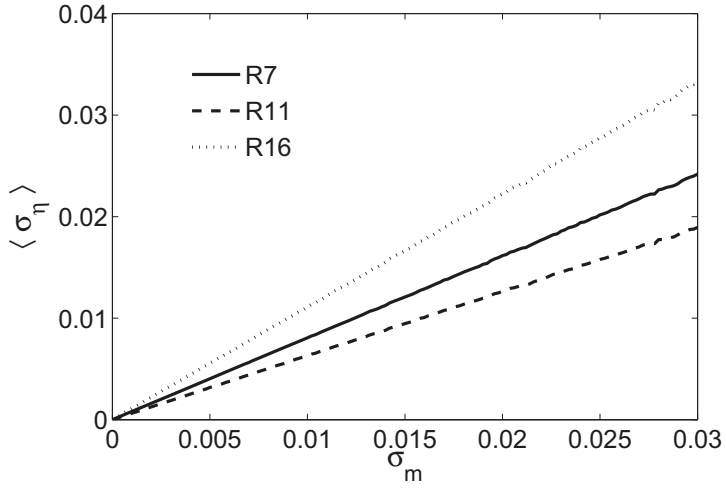

Figure 4.11: Effect of the photometric zero-points on the spread of overdensities η in cells. For each σ_m, 300 sets of $\Delta\boldsymbol{m}$ variates are generated and applied to each cell through $\boldsymbol{\eta} = \mathbf{A}\cdot\Delta\boldsymbol{m}$. The standard deviation of $\boldsymbol{\eta}$ across all cells is found for each set and averaged. Those averages are reported in the figure with σ_m in increments of 0.0002. The uncertainty in the reported averages is at the 1% level.

$$\mathbf{U} = \begin{bmatrix} | & & | \\ \hat{\mathbf{u}}_1 & \cdots & \hat{\mathbf{u}}_N \\ | & & | \end{bmatrix} . \tag{4.22}$$

Hereafter, we will refer to the orthonormal set of basis vectors contained in $\mathbf{U}$ as the *noise basis* which spans *noise-space*. Associated eigenvalue/eigenvector pairs will be referred to as the *noise eigenmodes*. Noise eigenmodes are indexed according to their eigenvalues from largest ($i = 1$) to smallest ($i = N$).

Vectors are rotated between noise-space and cell-space through $\mathbf{t} = \mathbf{U}^T\boldsymbol{\eta}$ and $\boldsymbol{\eta} = \mathbf{U}\mathbf{t}$. The columns of $\mathbf{U}$ comprise a full basis that spans cell-space. Therefore, any zero-point noise vector $\boldsymbol{\eta}$ can be represented as a linear combination of them,

$$\boldsymbol{\eta} = \sum_{i=1}^{N} \mathfrak{t}_i \hat{\boldsymbol{u}}_i, \tag{4.23}$$

where the expansion coefficient $\mathfrak{t}_i$ equals the projection of the zero-point noise onto the i^{th} noise mode, $\mathfrak{t}_i = \boldsymbol{\eta} \cdot \hat{\boldsymbol{u}}_i$.

A random zero-point overdensity vector $\boldsymbol{\eta}^{(\tau)}$ in cell-space may be generated by populating a 2052-element long vector $\Delta\boldsymbol{m}^{(\tau)}$ with i.i.d. Gaussian offsets and rotating through $\mathbf{A}$, i.e. $\boldsymbol{\eta}^{(\tau)} = \mathbf{A} \cdot \Delta\boldsymbol{m}^{(\tau)}$. Zero-point noise realizations can also be generated, albeit in a less efficient way, by utilizing the statistical orthogonality of $\mathfrak{t}_i$ and applying noise-space versions of equations (4.7) and (4.8).

There are only 2052 PRIMARY SEGMENTs, so the zero-point noise can have at most 2052 degrees of freedom. In practice, this ends up being an overestimate since many regions have areas close to or equal to zero.

The actual dimensionality of the noise is revealed through an inspection of its eigenvalues. In Figure 4.12 we see that $\lambda_i^{(\eta)} \approx 0 \; \forall \, i > 1890$. This indicates that 1890 PRIMARY SEGMENTs, at most, have any appreciable impact on the noise. Figure 4.12 also shows that the R11 and R16 cells have several more "nonzero modes" than do the R7 cells. This is likely a result R11/R16's larger angular radii, which permit their cells to extend further beyond the boundaries of the spectroscopic footprint where additional PRIMARY SEGMENTs can be "picked up".

Another perspective of the noise structure is provided in Figure 4.13. We see here that 90% of the noise variance is contained within the first 339, 217, and 202 modes for R7,

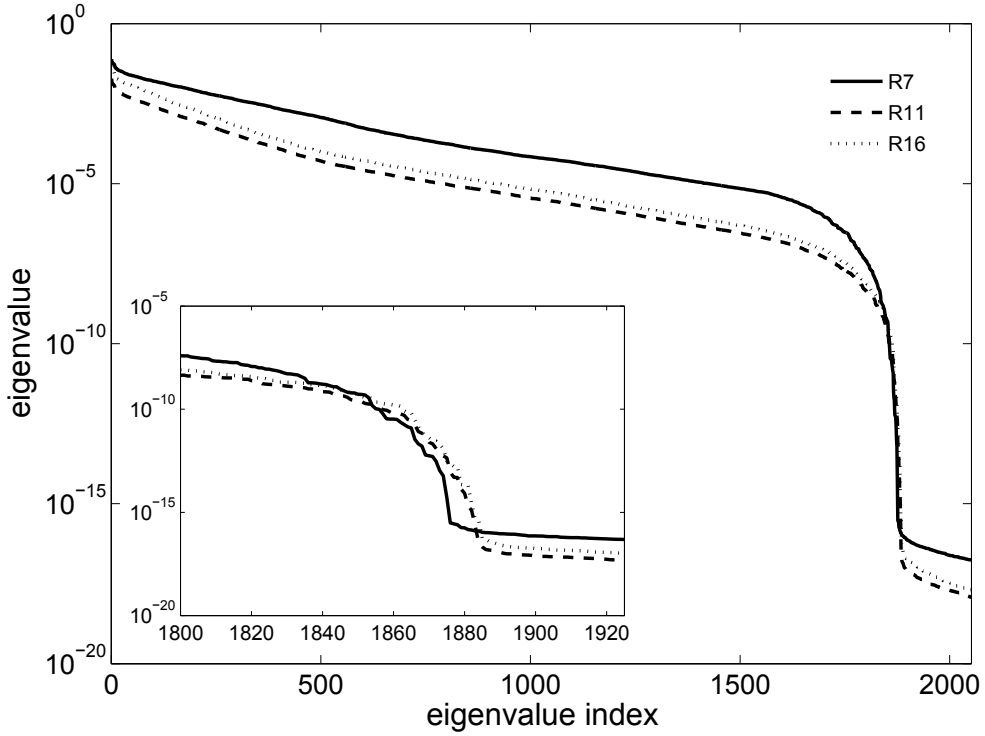

Figure 4.12: Scree plot of zero-point eigenvalues $\lambda^{(\eta)}$ when $\sigma_m = 0.01$. Eigenvalues are ranked from largest (index = 1) to smallest (index = N).

R11, and R16. The zero-point noise can be mostly represented with just 1% of the modes available in cell-space. This stands in stark contrast to the signal, for which 78-89% of the modes were required to recover the same percentage of the variance.

The total zero-point noise variance equals $\sum_{i=1}^{N} \lambda_i^{(\eta)} = \mathrm{tr}\big(\mathbf{\Lambda}^{(\eta)}\big)$. We find,

$$\sum \lambda_{R7}^{(\eta)} = 5.16, \quad \sum \lambda_{R11}^{(\eta)} = 0.79, \quad \sum \lambda_{R16}^{(\eta)} = 1.89. \tag{4.24}$$

Per cell this works out to be 6.54×10^{-5}, 4.03×10^{-5}, and 12.47×10^{-5} for R7, R11, and R16. This ordering is consistent with that shown in Figure 4.11.

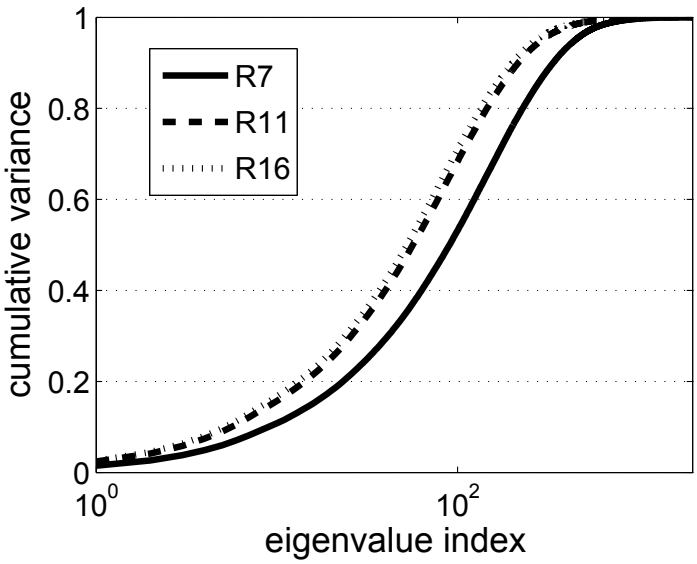

Figure 4.13: Cumulative variance of zero-point noise eigenmodes. For each eigenvalue index n, the variance reported on the vertical axis is $\sum_{i=1}^{N} \lambda_i^{(\eta)}$. This and Figure 4.12 reveal that the R7 modes have a more equal distribution of variance between them. The R11 and R16 noise modes are more front-loaded, capturing a greater percentage of the noise variance in the early modes, with a sharper diminishment thereafter.

Spatial depictions of four R7 noise eigenvectors are presented in Figure 4.14. Modes 1 and 2 are three-dimensional representations of the longest PRIMARY SEGMENTs in the northern and southern skies, respectively. The magnitudes of eigenvector elements tend to increase with redshift since zero-points mostly impact η_i in distant cells. Higher order noise modes are preferentially linear combinations of shorter regions. They become increasingly diffuse towards the degrees-of-freedom limit. For a slideshow of more zero-point noise eigenmodes, visit this link.

The frequency compositions of the noise eigenvectors are revealed through their power spectra. A collection of these are provided in Figure 4.15. While the signal spectra are approximately smoothed Dirac delta functions, the noise spectra have more equal power

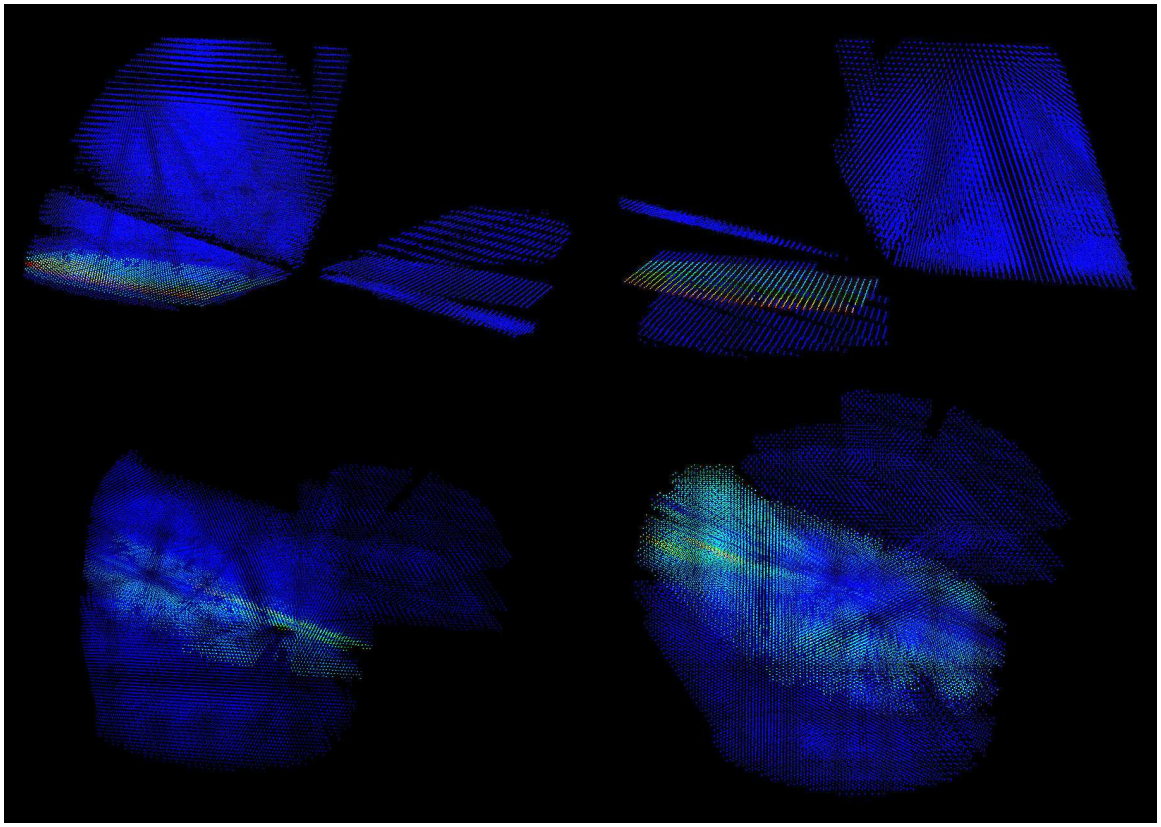

Figure 4.14: Visual depiction of four zero-point overdensity eigenvectors $\hat{\boldsymbol{u}}$. Each pixel represents one R7 cell. Color stands for the magnitude of the eigenvector element in each cell with dark blue being of lowest magnitude and red being of highest magnitude. Starting in the upper-left corner and running clockwise these are modes 1, 2, 100 and 25.

across frequency space. Rather than fade away at low k, they tend to increase or stay constant. They are also more likely to possess local maxima in addition to their absolute peaks.

Taken together, these features reveal that the noise eigenmodes possess a richer tapestry of structure than do the signal modes. This is because PRIMARY SEGMENTs have distinct features along three separate dimensions.

The first is depth. While PRIMARY SEGMENTs are fundamentally two-dimensional

areas, their physical manifestations with respect to zero-point errors are three-dimensional slices in redshift space. These slices span the full redshift range of the cells. For R7/R11 and R16, these amounts to $\log_{10}(k) = -1.95$ and $-2.09\,h\,\mathrm{Mpc}^{-1}$, respectively, and helps explain the prevalence of power on large scales.

The second dimension is the height of the SEGMENTs. Each is 10 to 12 arcminutes tall. Depending on whether these are assessed at the near or far end of the survey, they account for structures on scales between $\log_{10}(k) = 0.37$ and $1.51\,h\,\mathrm{Mpc}^{-1}$. These are well beyond the resolution of both the cells and FFT grid, and do not impact these spectra.

The final dimension is the length of the PRIMARY SEGMENTs, the longest of which is about $130°$. As Figure 4.14 and associated slideshow illustrate, a noise eigenmode may contain a linear combination of modes of different lengths, though there is a tendency for PRIMARY SEGMENTs of similar length to "flock together." These are the structures that leave their imprint in the form of absolute and local maxima, and which provide each noise spectrum its unique character.

It is still possible to determine a single characteristic scale for each noise eigenmode by identifying the wavenumber at which it achieves peak power. We do so in Figure 4.16, but caution that these results must be interpreted more carefully than those for the signal, for the reasons just mentioned. Perhaps the most important takeaway is that almost all noise modes of any significance lie squarely in the linear regime.

The noise eigenmodes also tend to overlap one another in Fourier space. A video simulation of selected modes can be viewed at this link. The most obvious feature is that the

modes are largely spread out along the xz-plane, with contributions from only the smallest k_y values. This fact is perhaps best explained by Figure 4.17, which shows that the DR6 PRIMARYs are oriented along the y-axis. The structures along this axis are preferentially long and of low frequency.

This investigation into signal and noise structure is done in pursuit of a method to reduce systematic noise in measurements of overdensities. We learned that while each signal mode represents its own frequency, noise modes tend to overlap and have power across the spectrum. This latter feature is troubling. Because noise modes cannot be easily isolated in frequency space, no simple bandpass filter can remove them.

This ultimately suggests a flaw in the SDSS survey design. Having STRIPEs of different lengths makes zero-point effects harder to localize. If the STRIPEs were more uniform, or if the survey was taken in equally-sized patches, noise could be addressed in a more targeted manner. Instead, the SDSS strategy has introduced systematic errors on all length scales and in two of three dimensions.

Another line of attack in noise reduction occurs in configuration-space. The fact that zero-point noise $\boldsymbol{\eta}$ — originally represented in tens of thousands of dimensions across all cells — can be localized as $\mathbf{t}$ in just a couple hundred dimensions offers an attractive opportunity. It suggests the possibility of targeting the lowest order noise eigenmodes without substantially impacting the signal. One idea is to deproject all the data (e.g. signal *and* noise) that lies along some set of $\hat{\boldsymbol{u}}_i$ in an effort to maximize the signal-to-noise ratio.

It turns out that the most important signal modes preferentially overlap the most prin-

ciple noise modes, a conclusion that might have been drawn from a comparison of Figures 4.8 and 4.16. Removing information along the latter inevitably removes much of the former. Under certain conditions, e.g. high magnitude noise, this might be worth it. But in this case, simple deprojection is the equivalent of cutting off your finger to remove dirt under the nail.

These conclusions were arrived at only after great toil. Many variations of noise reduction techniques that exploit the limited noise dimensionality were theorized, developed, tested, and ultimately discarded for use in this application. However, in other contexts we believe these solutions still hold promise. For this reason, and in an effort to ensure this work doesn't go to waste, we report these theories in Appendix G.

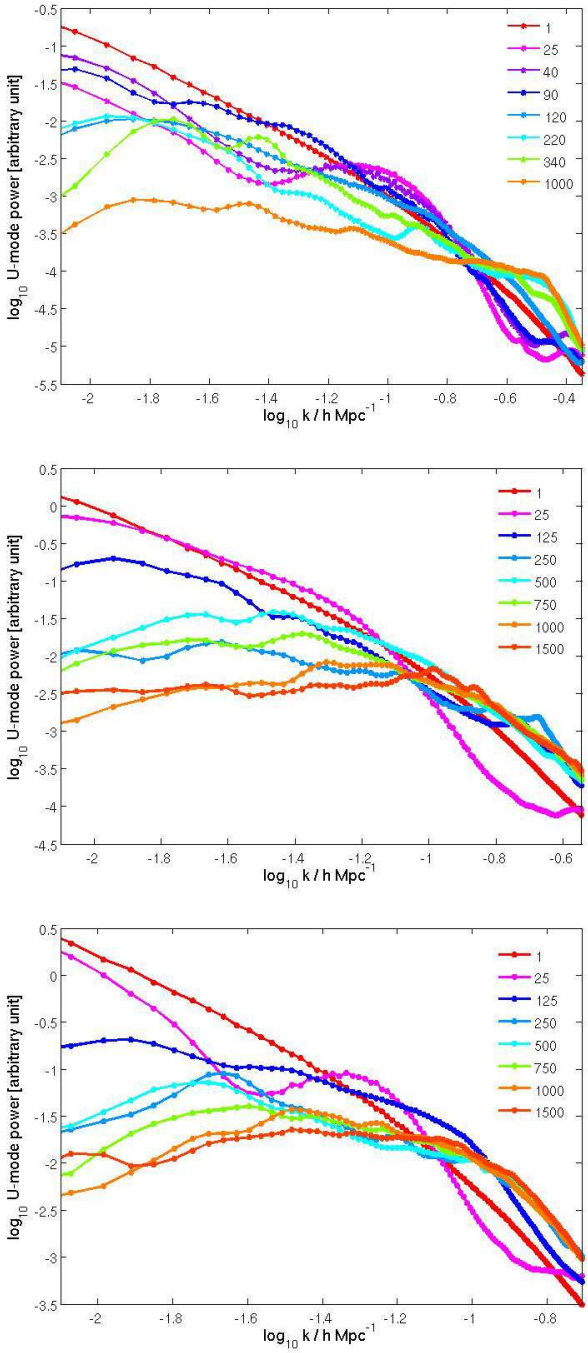

Figure 4.15: Power spectra of select zero-point noise eigenvectors $\hat{\boldsymbol{u}}_i$ for R7 (*top*), R11 (*middle*) and R16 (*bottom*).

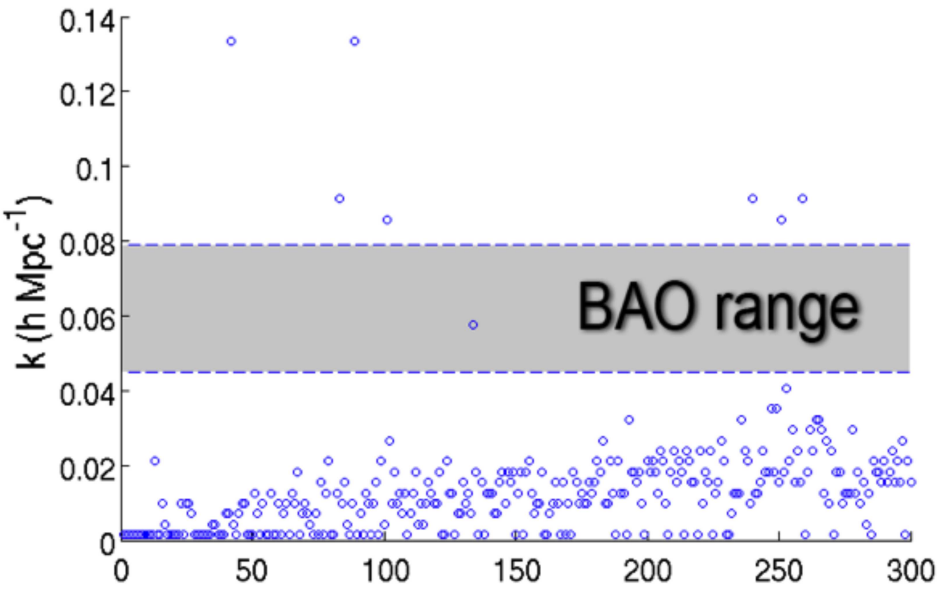

Figure 4.16: Wavenumbers principally represented by each of the first 300 R7 zero-point noise eigenmodes. The horizontal axis contains the indices of the eigenmodes. The values on the vertical axis are the peaks of each noise mode's power spectrum. Wavenumbers appear quantized due to the finite number of k-bins. The typical range of the BAOs, $0.045 \leq k \leq 0.079\,h\,\mathrm{Mpc}^{-1}$, is shaded in gray.

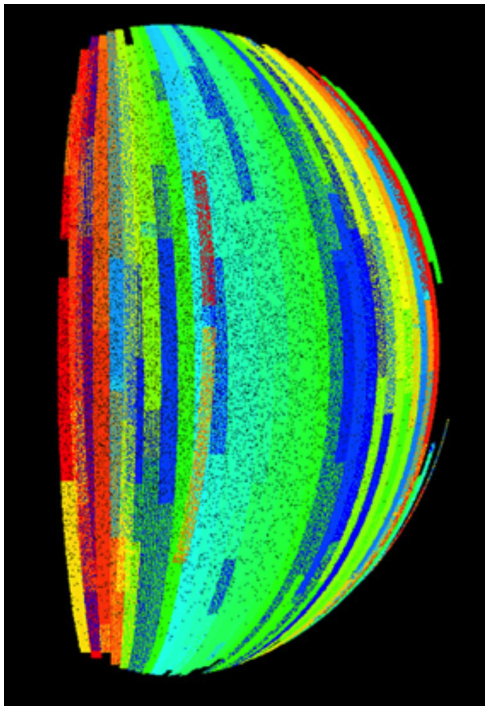

Figure 4.17: Northern cap of the DR6 PRIMARY footprint with the y-axis aligned vertically. PRIMARYs are assigned random colors to distinguish them from their neighbors. At a given redshift, photometric offsets affect number counts in PRIMARY SEGMENTs uniformly. With the noise eigenmodes this leads to large scale structures in y and a suppression of structure at high k_y.

4.7 Notation Summary

The amount of notation needed to keep track of all the signal and noise components in different spaces is daunting. As a reference, here we include all notation needed going forward. At times, we will have little alternative but to overload these variables in describing one problem or another. In such cases, we will make it clear that the variable's definition should be limited to the current context and not be applied universally. As a general rule, vectors will be bolded and italicized while matrices will be bolded, capitalized and not italicized.

A vector whose i^{th} element equals the value in the i^{th} spherical cell is said to be represented in *cell-space*. There are four other coordinate systems defined and utilized within this thesis. These are *signal-space*, *noise-space*, W*-space* and B*-space*. The symbols used for each component in each space are provided in Table 4.3.

Space	Data	Signal	Zero-Point Noise	Shot Noise
cell	$\boldsymbol{\Gamma}$	$\boldsymbol{\kappa}$	$\boldsymbol{\eta}$	$\boldsymbol{\zeta}$
signal	$\boldsymbol{M}$	$\boldsymbol{S}$	$\boldsymbol{T}$	$\boldsymbol{Y}$
noise	$\mathbf{m}$	$\mathbf{s}$	$\mathbf{t}$	$\mathbf{y}$
W	$\boldsymbol{m}$	$\boldsymbol{s}$	$\boldsymbol{t}$	$\boldsymbol{y}$
B	$\boldsymbol{\xi}$	$\boldsymbol{\omega}$	$\boldsymbol{\phi}$	$\boldsymbol{\pi}$

Table 4.3: Notation used to represent data, signal, shot noise, and zero-point noise in the five bases (or spaces) considered.

Regardless of which space one chooses to represent their data, we can always define a basis-independent set of parameters $\boldsymbol{p} = \{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\varphi}\}$ that uniquely describes any overdensity vector. These are given in Table 4.4. For example, in cell-space $\boldsymbol{\Gamma}(\boldsymbol{p}) = \boldsymbol{\kappa}(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\psi}) +$

$\boldsymbol{\zeta}(\boldsymbol{\varphi})$. In signal-space $\boldsymbol{M}(\boldsymbol{p}) = \boldsymbol{S}(\boldsymbol{\theta}) + \boldsymbol{T}(\boldsymbol{\psi}) + \boldsymbol{Y}(\boldsymbol{\varphi})$, and so on. For N-dimensional vectors, a maximum of N parameters are needed. If the rank of one's signal, noise or data is $< N$, then only as many parameters as there are degrees of freedom are required.

Data	Signal	Zero-Point Noise	Shot Noise
$\boldsymbol{p}$	$\boldsymbol{\theta}$	$\boldsymbol{\psi}$	$\boldsymbol{\varphi}$

Table 4.4: Basis-independent vectors used to parameterize data, signal, shot noise, and zero-point noise.

When random processes are simulated for Monte Carlo purposes, each random vector will receive a superscript to indicate which realization it is. The general index for realizations is τ. For instance, the τ^{th} simulated data vector in cell-space would be

$$\boldsymbol{\Gamma}^{(\tau)} = \boldsymbol{\kappa}^{(\tau)} + \boldsymbol{\eta}^{(\tau)} + \boldsymbol{\zeta}^{(\tau)}. \tag{4.25}$$

The covariance matrix of the overdensities in cell-space is denoted by $\boldsymbol{\Sigma}$. It is the sum of the covariance matrices of the signal $\boldsymbol{\Sigma}_\kappa$, the systematic noise $\boldsymbol{\Sigma}_\eta$, and the shot noise $\boldsymbol{\Sigma}_\zeta$ such that

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\kappa + \boldsymbol{\Sigma}_\eta + \boldsymbol{\Sigma}_\zeta. \tag{4.26}$$

When the covariance matrices of two processes need to be manipulated as a unit, a double subscript signals a matrix sum,

$$\boldsymbol{\Sigma}_{\eta\zeta} = \boldsymbol{\Sigma}_\eta + \boldsymbol{\Sigma}_\zeta, \tag{4.27}$$

Space	i^{th} eigen-vector	i^{th} eigen-value	Eigenvector Matrix	Eigenvalue Matrix	Diagonalization
cell	$\hat{\boldsymbol{e}}_i$	1	$\mathbf{I}$	$\mathbf{I}$	N/A
signal	$\hat{\boldsymbol{z}}_i$	$\lambda_i^{(\kappa)}$	$\mathbf{Z}$	$\mathbf{\Lambda}^{(\kappa)}$	$\mathbf{\Sigma}_\kappa = \mathbf{Z}\mathbf{\Lambda}^{(\kappa)}\mathbf{Z}^T$
noise	$\hat{\boldsymbol{u}}_i$	$\lambda_i^{(\eta)}$	$\mathbf{U}$	$\mathbf{\Lambda}^{(\eta)}$	$\mathbf{\Sigma}_\eta = \mathbf{U}\mathbf{\Lambda}^{(\eta)}\mathbf{U}^T$
W	$\hat{\boldsymbol{w}}_i$	$\lambda_i^{(W)}$	$\mathbf{W}$	$\mathbf{\Lambda}^{(W)}$	$\mathbf{\Sigma}_\kappa^{-1} + \mathbf{\Sigma}_{\eta\zeta}^{-1} = \mathbf{W}\mathbf{\Lambda}^{(W)}\mathbf{W}^T$
B	$\hat{\boldsymbol{b}}_i$	$\lambda_i^{(B)}$	$\mathbf{B}$	$\mathbf{\Lambda}^{(B)}$	$\mathbf{\Sigma}_\zeta^{-1} + \mathbf{\Sigma}_{\kappa\eta}^{-1} = \mathbf{B}\mathbf{\Lambda}^{(B)}\mathbf{B}^T$

Table 4.5: Summary of the notation used to represent each of five coordinate systems invoked in this thesis. Each eigenvector is a linear combination of cell-space eigenvectors $\hat{e}$. Eigenvectors stored successively in columns form eigenvector matrices. The products of eigenvector and eigenvalue matrices produce the covariance matrices indicated in the final column.

$$\mathbf{\Sigma}_{\kappa\eta} = \mathbf{\Sigma}_\kappa + \mathbf{\Sigma}_\eta. \tag{4.28}$$

With the exception of cell-space, every space is defined by the diagonalization of a covariance matrix. Table 4.5 summarizes the transformations and eigenmodes. Eigenvectors are ordered according to their eigenvalues from largest ($i = 1$) to smallest ($i = N$). Eigenvector matrices have the i^{th} eigenvector stored in the i^{th} column. Eigenvalue matrices have the i^{th} eigenvalue in the i^{th} diagonal element and equal zero elsewhere.

While there can be an infinite number of random or simulated data vectors, there is only one true overdensity vector for the MGS. This overdensity in cell-space is denoted with $\boldsymbol{\delta}$. While its true signal, zero-point noise, and shot noise components are known only to nature, they are represented in the following way,

$$\boldsymbol{\delta} = \boldsymbol{\delta}_\kappa + \boldsymbol{\delta}_\eta + \boldsymbol{\delta}_\zeta. \tag{4.29}$$

	$\boldsymbol{\delta}$	$\boldsymbol{\kappa}$	$\boldsymbol{\zeta}$	$\boldsymbol{\eta}_{\sigma_m=0.01}$	$\boldsymbol{\eta}_{\sigma_m=0.02}$	$\boldsymbol{\Gamma}_{\sigma_m=0.01}$	$\boldsymbol{\Gamma}_{\sigma_m=0.02}$
R7	2.05101	1.13438	0.69651	0.000065	0.000260	1.83108	1.83101
R11	1.07395	0.58092	0.17032	0.000040	0.000160	0.75129	0.75128
R16	0.60285	0.30758	0.06351	0.000030	0.000120	0.37112	0.37137
R16 (all z)	2.21056	0.30759	1.07197	0.000124	0.000496	1.37980	1.37908

Table 4.6: Single-cell variances of simulated signal and noise components. The variance across all cells was calculated empirically for each of 10,000 realizations. These variances were averaged and reported in this table. The variances of the systematic zero-point noise and overall data vector $\boldsymbol{\Gamma}$ have been reported for two separate parameterizations of σ_m. To compare the three cell sets on equal footing, the 3$^{\rm rd}$ row contains the variances for R16 cells at $z < 0.22$, while the final row reports variances over all R16 cells.

4.8 Census of Simulated Variances

Throughout this chapter we have introduced methods by which signal and noise realizations can be generated from fiducial models. To quantify the importance of each component, we have reported the average of the single-cell variances in Table 4.6.

Chapter 5

Photometric and Spectroscopic Footprint Corrections

Well-defined galaxy survey footprints are a necessity for precision cosmology. Accurate boundaries limit our focus to regions where data has been collected and processed in a consistent way. A footprint's area can define the average angular density of objects, a necessary statistic for calculating the expected numbers of galaxies in cells. We use them to determine spectroscopic completeness along lines-of-sight and to decide how to account for targets without spectra. They help us to better understand geometry-dependent effects like zero-point photometric offsets, and much more.

As first discussed in §2.2, there are two SDSS footprints that hold the greatest importance. The first is the *photometric footprint*, which we define to be the union of PRIMARY SEGMENTs. Recall that PRIMARY SEGMENTs contain primary photometric observa-

tions of galaxies, are strictly nonoverlapping, and are largely contiguous. While there are photometrically-detected objects outside the union of PRIMARY SEGMENTs, none of these are considered in any fashion in this analysis.

The second footprint of importance is the *spectroscopic footprint*, which we define to be the union of SECTORs. SECTORs result from the intersections of spectroscopic TILEs and masks. They may be thought of as the smallest structures within the spectroscopic footprint; even adjacent SECTORs can have different completeness properties. Only targets within SECTORs have redshifts, making the spectroscopic footprint the most vital area for creating accurate three-dimensional galaxy maps.

As the SDSS progressed, new TILEs were periodically placed. The intersections between them and existing TILEs created hundreds of new SECTORs at a time, many obtaining new spectroscopic properties. In this way, the spectroscopic footprint was complex and in transition. Tiling runs for DR7 were planned during DR6. This "forward-looking" approach facilitated the continuous evolution of the survey, but also introduced inconsistencies at the times of data releases, including that of DR6.

For example, consider a DR7 TILE partially overlapping the southern portion of a DR6 TILE. The tiling algorithm could restrict fiber placement to the northern portion of the DR6 TILE with the expectation that the remainder would be filled in during DR7. The entirety of the DR6 TILE would be included within the DR6 spectroscopic footprint even though the completeness was distinctly nonhomogeneous. This effect, which was largely limited to the edges of the spectroscopic footprint, could fool the user into believing that the DR6

footprint was larger than it actually is.

Assuming one was even aware of the existence of troublesome SECTORs, searches for them are not easy. Their region definitions are complicated and not well documented. Statistical tools to root them out might be developed, but settling on decision criteria is challenging. The TILEs within which they reside still receive their full compliments of fibers. Searching for empty regions alone is insufficient because troublesome SECTORs were routinely assigned a nonzero number of fibers (though certainly not enough to achieve representative spectroscopic completeness). Moreover, it is difficult to distinguish between regions that are systematically undersampled and those that are legitimately underdense.

We found that problems with the footprint definitions were detected more effectively by eye than through code. Using the VisIVO software package[1], we filtered millions of uniformly distributed angular randoms into sets that resided within particular regions. By superimposing positions of MGS galaxies and MGS objects, it was possible to locate the regions where the survey geometry did not match the observations. Specific examples of these discrepancies are provided in this chapter.

We will report on an area of the photometric footprint where PRIMARY SEGMENT definitions are ambiguous, making it difficult to know which galaxies are subject to which photometric offsets. We also discover five additional areas that ought to be removed from the photometric footprint. Failure to do so inflates cells' photometric footprint volumes

[1]The Visual Interface for the Virtual Observatory (VisIVO) was designed specifically to visualize large, astrophysical data sets. VisIVO visualizations use colored pixels to represent each point, allowing desktop computers to display hundreds of millions of elements at a time. Most depictions of the SDSS footprint and galaxies in this dissertation were generated through VisIVO.

β_{PS}, leading to underestimations of their overdensities and the introducion of artificial power spectrum features on the scales of those areas. We quantify how errors in β_{PS} affect the normalization of the selection function, estimates of the number of galaxies expected in cells, and all conclusions that follow.

In addition, we make corrections to the spectroscopic footprint by removing hundreds of troublesome SECTORs and restoring regions that were erroneously omitted. Without these improvements, it is impossible to guarantee that all cells that make it into the survey satisfy a common spectroscopic volume threshold β_{SPEC}.

Furthermore, as we will argue in Chapter 6, counting MGS targets in cells when they lack spectra poses a special challenge. We introduce a number of solutions but note that the preferred method depends upon the type of environment within which those objects reside. Accurate spectroscopic footprint definitions are critical to characterizing those environments. While the impacts listed above are those most relevant to our analysis, we note that there are certainly other measurements impacted by ill-defined photometric and spectroscopic footprints.

By the conclusion of this chapter we will have described the creations of an *improved photometric footprint* and an *improved spectroscopic footprint*. Together, they constitute what is almost certainly the most accurate geometric description of DR6 in existence.

5.1 Photometric Footprint

In this section, we discuss errors found in the DR6 photometric footprint and the steps taken to remedy them. We begin by studying one troublesome area in detail, showing both how the problem was discovered and how corrections are incorporated into the photometric footprint definitions. We continue by revealing problems in four other areas and in one STRIP. Finally, we quantify the effect these errors have on the expected numbers of galaxies in cells.

5.1.1 Locating and Correcting Footprint Problems

Figure 5.1 contains four circular TILEs superimposed upon a sea of MGS targets. A closer view of this area is shown in Figure 5.2. The areas boxed in red lie within the union of PRIMARY SEGMENTs, yet contain no MGS targets.

While the absence of MGS targets does not necessarily indicate a problem with the photometric footprint, there are three observations in this case that strongly suggest an error in the six PRIMARY SEGMENT definitions. First, given the ambient surface density of MGS targets, the probability that areas of this size would be empty due to cosmic variance alone is very low. Second, the shapes of the empty regions align perfectly with the SDSS geometry. Each boxed region corresponds to a single SEGMENT. These regions comprise a group of six within a single STRIP. Finally, these six areas are missing from the spectroscopic footprint, suggesting that whatever caused the lack of MGS targets was reflected in

the SECTOR definitions, but not the SEGMENT definitions.

Together, these observations provide compelling evidence that the six SEGMENTs in Figure 5.2 were included in the photometric footprint by mistake. We shall refer to their union as "Bad Area 1".

The footprint can be corrected if the constraint conditions that define the edges of Bad Area 1 can be identified. Then, points that lie within the union of PRIMARY SEGMENTs could be filtered through six additional searches over these SEGMENTs. Any points that lie inside any of those six regions would be summarily classified as residing outside the photometric footprint.

The constraint conditions for the edges of footprint errors are not always defined in the DR6 geometry. In our experience, and for reasons elaborated upon in §5.2, it is preferable to define the boundaries of troublesome regions using constraints reported in the DR7 region definitions.

The first step in defining Bad Area 1 is identifying the TILEs surrounding the six SEGMENT portions. Once approximate angular limits of the TILEs are found (lines of constant RA and declination can be generated by the user and superimposed onto one's visualization), a query such as the following can pick out the region IDs of the colored TILEs:

```
SELECT *
FROM BestDR7.dbo.fRegionsContainingPointEq(126,42.5,'TILE',0)
```

The goal is to discover the smallest set of TILEs whose interiors contain all boundaries of these six SEGMENTs. The user should pre-compute for each TILE a table that contains a list of its SECTORs and their definitions. This table should be organized such that one's

visualization software is able to display SECTORs individually, as in Figure 5.3. Here SECTOR 92487, colored in blue, is revealed to share the same boundaries as the bottom and right edges of the top SEGMENT in Bad Area 1.

The other boundary conditions are identified by examining SECTORs within TILE 2499. Figure 5.4 shows that the bottom edge of SECTOR 90846 is the same as the upper edge of the top SEGMENT. SECTOR 91521, pictured in Figure 5.5, possesses a comb-shaped border that traces every other boundary in Bad Area 1. Once the minimum number of SECTORs that share all of the SEGMENTs' boundaries are singled out, their constraint conditions are drawn from a pre-computed table of each SECTOR's halfspace constraints.

Recall that each constraint marks the intersection of a great or small circle with the unit sphere. Points lying on one side of the constraint occupy at least a full hemisphere and can be computationally expensive to search over. To speed up computations, the best method is to perform a preliminary filtering, perhaps by limiting points to those within a single TILE, and then applying the halfspace constraints one at a time.

Ultimately, 24 constraint conditions are needed to define Bad Area 1, or four for each of its six SEGMENTs. (Because these SEGMENTs share the same left and right boundaries, only 14 of these constraints are unique, however.) The aim is to report each SEGMENT's constraints such that any point that satisfies all four must lie within it.

Figure 5.6 illustrates one such constraint for the upper boundary of the fourth SEGMENT from the top. Points that satisfy this constraint are colored in blue and lie on the interior side of the SEGMENT, as desired. If the points had lain on the opposite side of the

boundary, all four constraint components $[n_x, n_y, n_z, c\,]$ would have been multiplied by -1 to flip the condition.

5.1.2 Census of Photometric Footprint Errors

The last section offered a detailed example of how to locate, identify and characterize Bad Area 1. In this section, we provide a census of four other "Bad Areas" and one "Bad STRIP" located in the SDSS's southern hemisphere.

The location of "Bad Area 2" is revealed in Figure 5.7. A zoomed-in version is shown in Figure 5.8. The combined area of these regions is sufficiently small that cosmic variance could plausibly explain the absence of MGS targets; however, these five rectangular areas are also missing from the union of SECTORs, so we find it more likely that they reflect a problem with the photometric footprint.

The location of "Bad Area 3" is revealed in Figure 5.9. A zoomed-in version is shown in Figure 5.10. Unlike Bad Areas 1 & 2, this area lies outside the spectroscopic footprint, making it impossible to use SECTOR constraints to define its shape. Instead, DR6 SEGMENT definitions are used to find the boundaries of the long edges (i.e. those that run roughly parallel to the lines of right ascension) with the exception of SEGMENT 1770's extreme edge, which is bounded by PRIMARY constraint condition 1225.

The lower declination side of Bad Area 3 is bounded by the edge of a DR6 PRIMARY given by constraint condition 1228. The database does not appear to contain any constraint for the opposite side, so it must be approximated by trial and error. This edge is roughly

parallel to condition 1228, so a modification of its c parameter is sufficient to shift the boundary. We found the appropriate constraint 4-vector to be

$$n_x = 0.122972102,$$

$$n_y = -0.720394564,$$

$$n_z = -0.682575662,$$

$$c = 0.021.$$

Bad Area 3 constitutes a genuine problem with the photometric footprint. In practice though, this region did not impact the galaxy density analysis since no cells were placed in its vicinity.

The location of "Bad Area 4" is circled in Figure 5.11 and shown in detail in Figure 5.12. The location of "Bad Area 5" is shown in Figure 5.13. Both were identified relatively easily by finding areas without MGS targets that coincided with holes in the spectroscopic footprint.

In summary, the geometric descriptions of Bad Areas 1-5 are fully provided by the constraint conditions present in the following DR7 SECTORs:

Area 1: 90846, 92485, 92487

Area 2: 91430, 91439, 91440, 92430, 92438

Area 3: N/A

Area 4: 92220, 92658, 92673

Area 5: 85193, 85549, 85560, 85563, 85642, 85654

We refer to the final area of the photometric footprint that requires correction as the "Bad STRIP". It is pictured in Figure 5.14. PRIMARY SEGMENTs are defined to be strictly nonoverlapping, but we see here that SEGMENTS 5344-5349 and SEGMENTs 6874-6879 do overlap at the edge of PRIMARY 308.

The STRIP complimentary to SEGMENTs 6874-6879 does contain MGS galaxies, yet is undefined in both the photometric and spectroscopic region geometries. This suggests that its omissions from the SECTOR and SEGMENT definitions are errors. There are a couple possible explanations. The first is that the existence of SEGMENTs 6874-6879 is a mistake, meaning that SEGMENTs 5344-5349 are defined correctly, but its complimentary STRIP either doesn't extend far enough (i.e. all the way to the next PRIMARY) or it was legitimately truncated early. Another possibility is that SEGMENTs 6874-6879 are real and SEGMENTs 5344-5349 extend too far beyond their true boundary.

Either way, this introduces significant ambiguity regarding what is actually happening in this area. If one's research depends intimately the PRIMARY SEGMENT within which an object lies, as the photometric zero-points do, this mangling of region definitions complicates efforts to handle these MGS targets with fidelity.

One conservative solution would be to consider SEGMENTs 6874-6879 as real and manufacture its six complimentary SEGMENTs. While these 12 SEGMENTs might truly belong on the edge of PRIMARY 308, splitting them off would merely reduce the statistical knowledge that could be gained by knowing that the measured magnitudes in the adjacent

	Area in deg^2	**Percentage of Original PRIMARY SEGMENT Footprint**	**Percentage of Total Sphere**
Original PRIMARY SEGMENT Footprint	8303.96	100.000	20.1308
Bad Area 1	1.9805	0.0239	0.00480
Bad Area 2	0.5084	0.0061	0.00123
Bad Area 3	0.7008	0.0084	0.00170
Bad Area 4	1.3334	0.0161	0.00323
Bad Area 5	10.3819	0.1250	0.02517
Bad STRIP	4.8419	0.0583	0.01174

Table 5.1: The size of the Bad STRIP and each of the 5 Bad Areas identified in the photometric footprint. Values were derived empirically using 3.85×10^8 full sky angular randoms filtered through each region.

regions are correlated.

However, our solution in this analysis was to remove the regions covered by SEGMENTs 6874-6879 from the photometric footprint entirely. This slightly reduces its area but also removes any worry that the objects in this region should have been excluded for a legitimate reason. The contiguous area covered by this portion PRIMARY 308 is defined by constraint conditions 8466, 15572, 15573 and 15574.

Table 5.1 summarizes the sizes of each region removed from the photometric footprint. The sum of all five Bad Areas plus the Bad STRIP is 19.7474 deg^2 . This is 0.237807% of the original PRIMARY SEGMENT footprint and 0.047872% of the total sphere. In the end, the area of the improved photometric footprint becomes 8284.21 deg^2 , a reduction of about 133 deg^2 or 1.58% from the reported value of 8417 deg^2 (Adelman-McCarthy et al., 2008).

5.1.3 Impact on Expected Number Count

By assuming that galaxies can exist in Bad Areas, $\langle n \rangle$ within cells that intersect those areas will be overestimated. In turn, δ will be underestimated. This type of error is bound to be most pronounced in high-redshift cells whose smaller angular radii are more likely to be occulted by Bad Areas. However, low-redshift cells can potentially intersect multiple PRIMARY SEGMENTs in the same Bad Area so the effect is worth examining empirically.

We begin by letting $\beta_{PS}^{(0)}$ equal the fraction of a cell's volume that lies within the union of PRIMARY SEGMENTs and β_{PS} equal the fraction of a cell's volume within the improved photometric footprint. The number of expected galaxies will be overestimated by the fraction $\left(\beta_{PS}^{(0)} - \beta_{PS}\right)/\beta_{PS}$.

All cells that intersect Bad Areas are identified, and using the Monte Carlo method outlined in Appendix E, their $\beta_{PS}^{(0)}$ and β_{PS} and values are calculated. We present the fractional overestimates of $\langle n \rangle$ in Figure 5.15. Because no cells intersect Bad Area 3, it is omitted from the Figure. The Bad STRIP is likewise omitted since its problem is not an absence of expected galaxies, but rather an ambiguity regarding PRIMARY SEGMENT definitions.

The error in $\langle n \rangle$ can be significant, exceeding 60% for some R7 cells. The maximum possible error tends to increase with redshift since the areas of the most distant cells decrease while the angular extent of the Bad Areas remains fixed. Due to its larger area, the errors are greatest in Bad Area 5. By virtue of their size, R16 cells have smaller fractions of their volumes affected by Bad Areas. It follows that their average fractional error in $\langle n \rangle$

is smaller than that of the R7 cells.

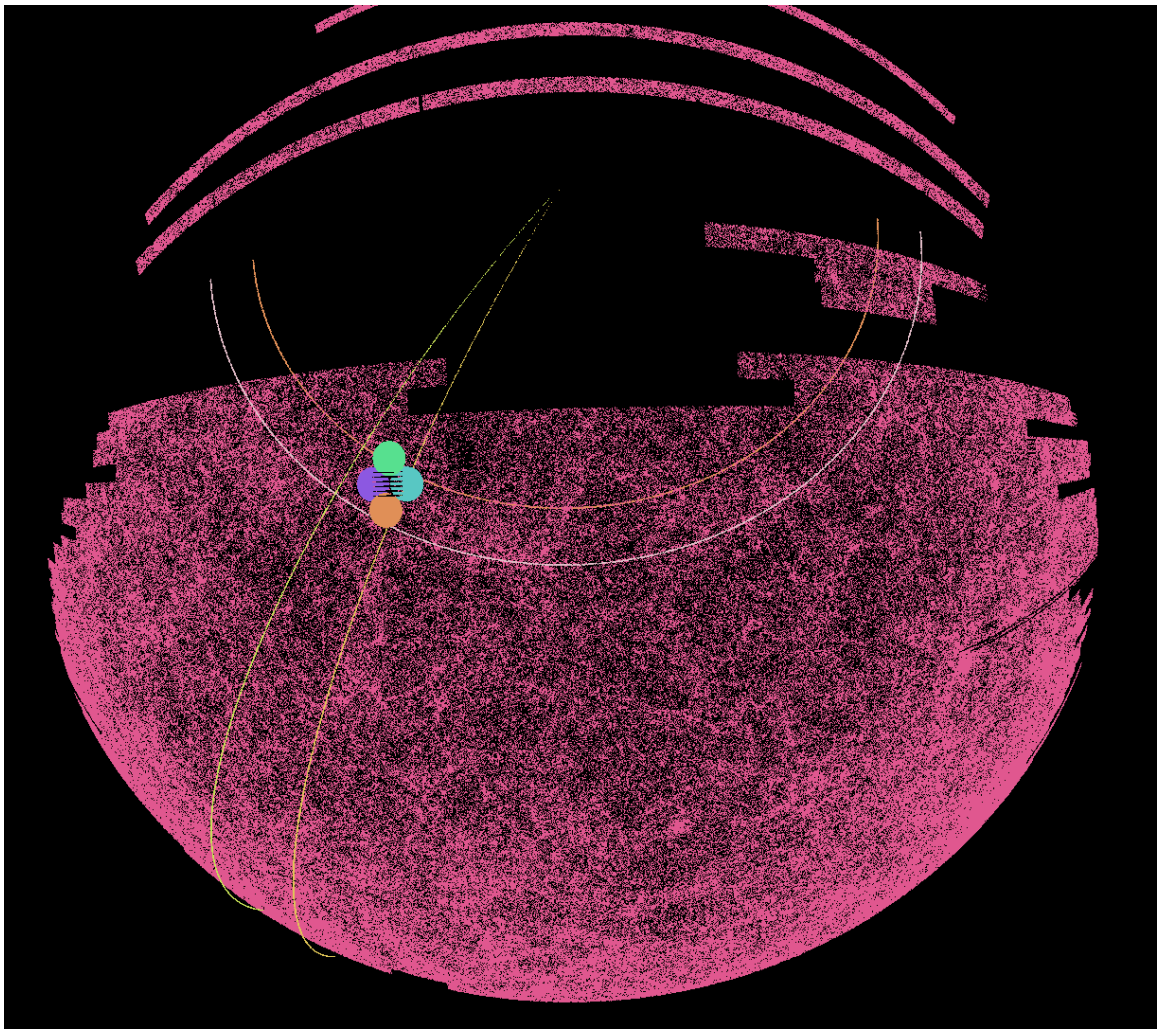

Figure 5.1: Wide view of Bad Area 1. Starting at the top and going clockwise, the colored circles are DR7 TILEs 2660, 2500, 2818 and 2499. They are superimposed atop MGS targets, represented by pink pixels, that indicate the extent of the DR6 PRIMARY SEGMENT footprint. The curved lines mark R.A.'s of 145° and 155° (left to right) and declinations of 55° and 60° (bottom to top).

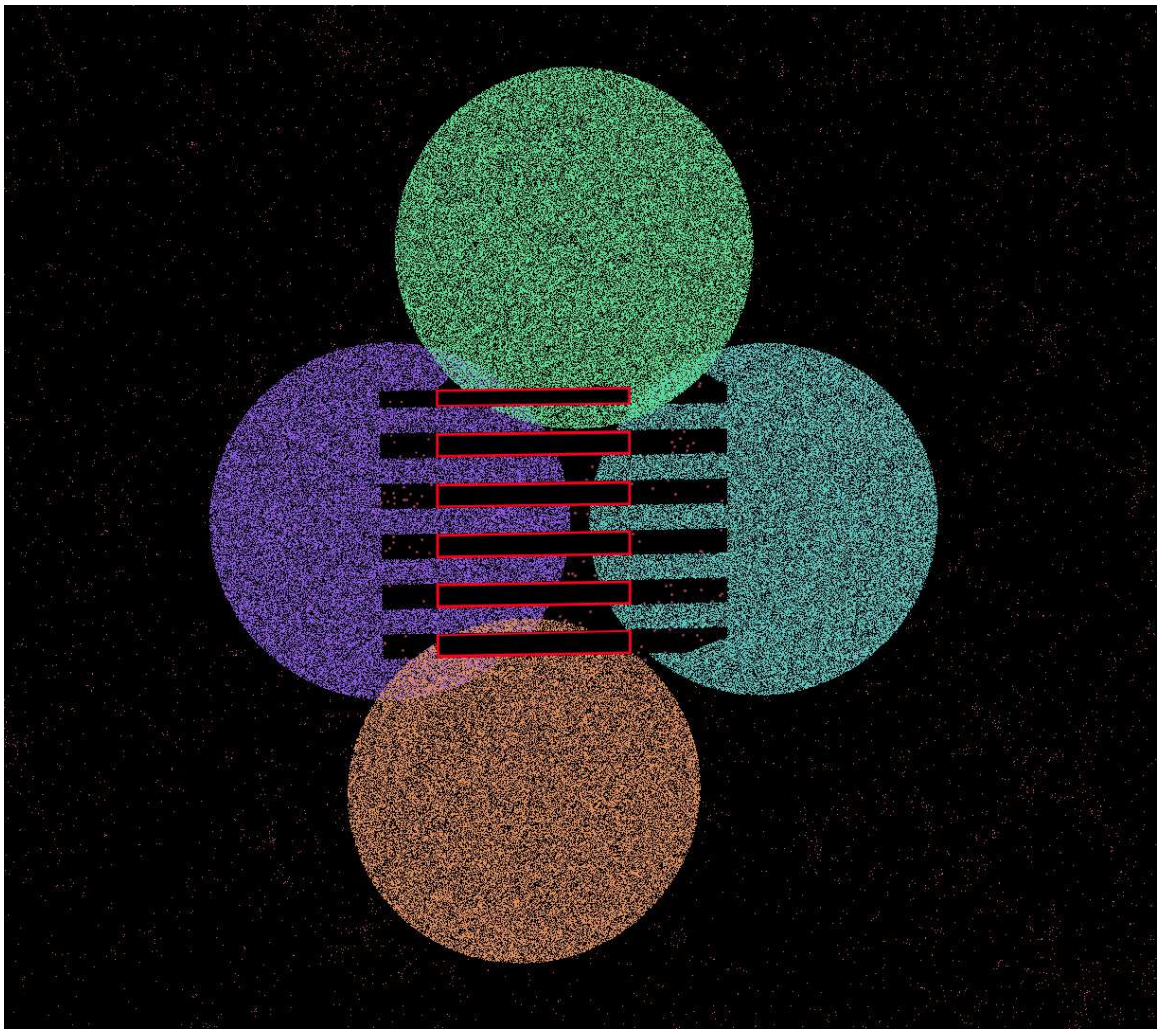

Figure 5.2: Close view of Bad Area 1. The colored circles contain the portions of DR7 TILEs 2660, 2500, 2818 and 2499 that lie within the DR6 spectroscopic footprint. The positions of MGS targets are marked by pink pixels. The areas boxed in red are the portions of six SEGMENTs within the PRIMARY SEGMENT footprint that contain no MGS targets.

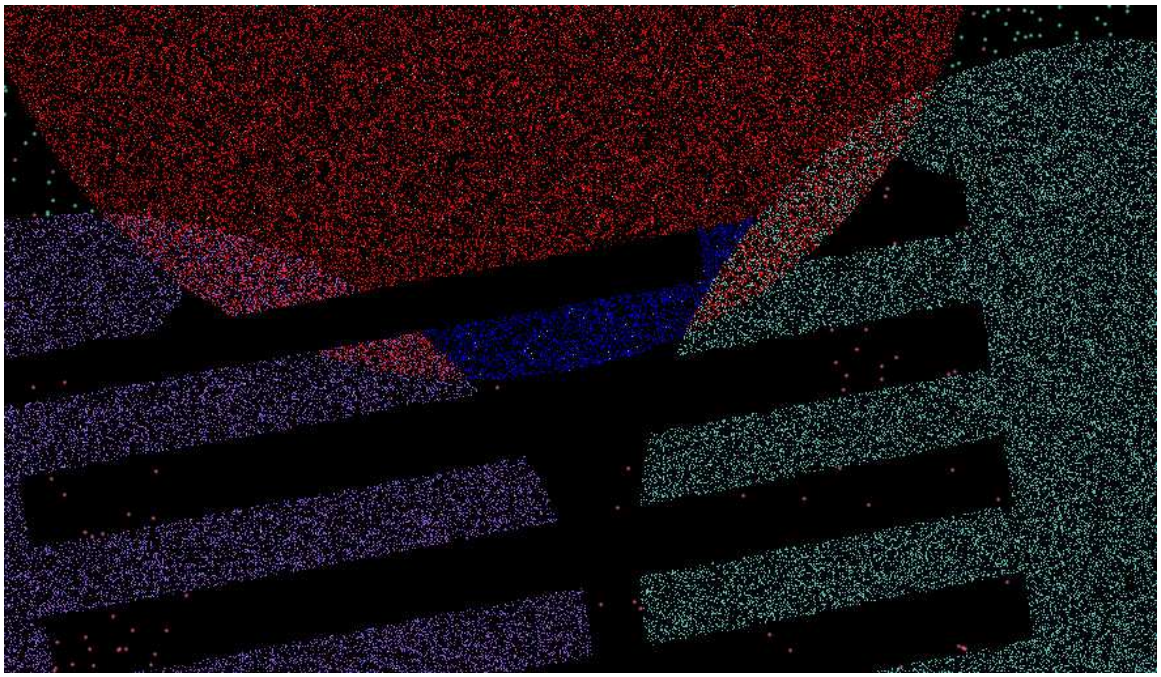

Figure 5.3: DR7 SECTOR 92487, colored in blue, lies within the area of DR7 TILE 2660, colored in red. MGS objects and MGS targets are represented by pink and green pixels, respectively. The union of SECTORs belonging to DR7 TILEs 2499 and 2500 are colored in purple and aquamarine, respectively.

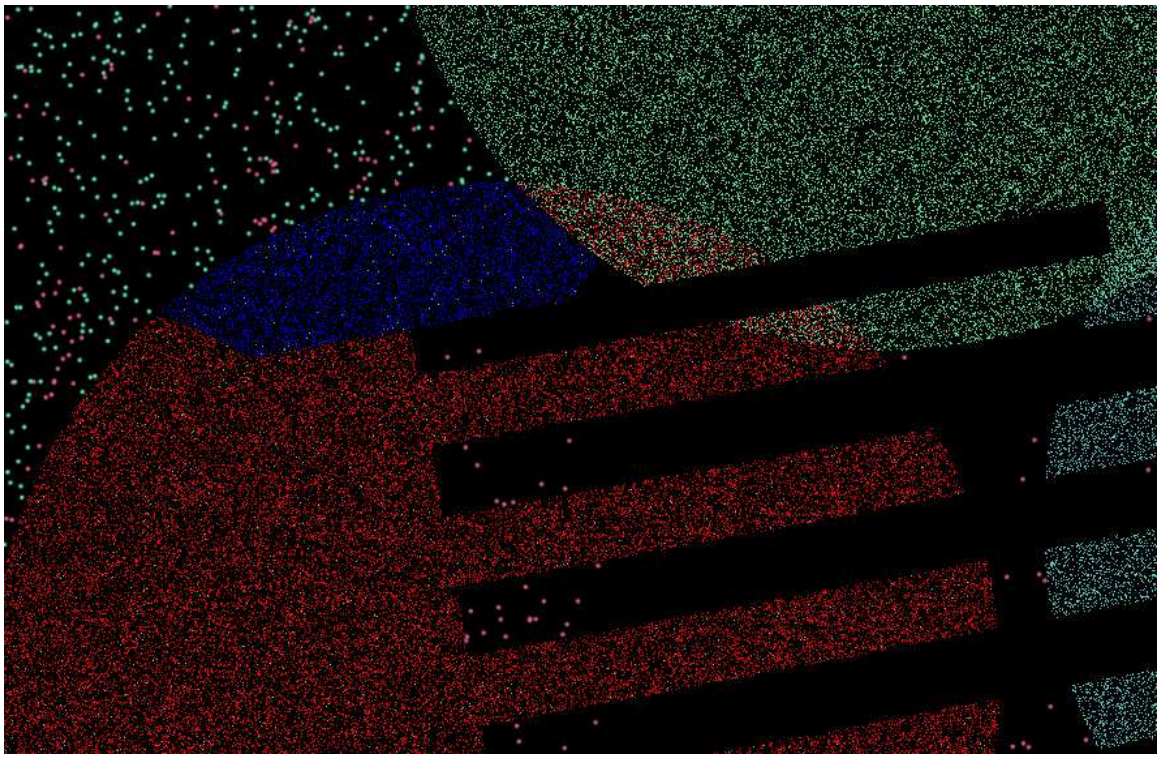

Figure 5.4: DR7 SECTOR 90846, colored in blue, shares its lower boundary with the upper boundary of the top SEGMENT in Bad Area 1. This SECTOR is one of the set belonging to DR7 TILE 2499, colored in red. MGS galaxies are represented by medium aquamarine pixels while MGS objects are represented by dark pink pixels. DR7 TILEs 2600 (*green*) and 2500 (*cyan*) are also pictured.

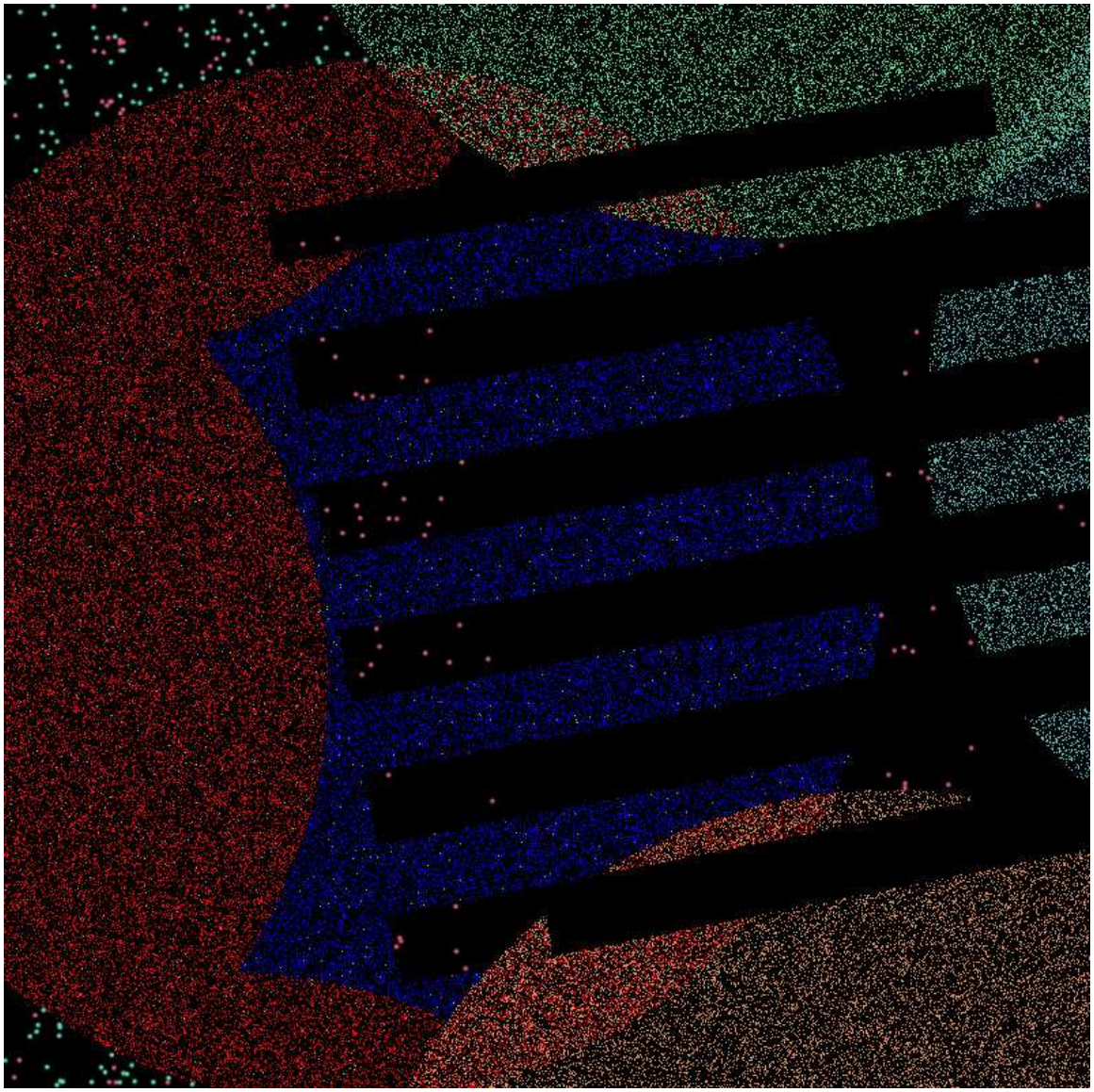

Figure 5.5: DR7 SECTOR 91521, colored in blue, shares boundaries with all six SEGMENTs in Bad Area 1. This SECTOR lies within DR7 TILE 2499, colored in red. MGS galaxies are represented by medium aquamarine pixels while MGS objects are represented by dark pink pixels. DR7 TILEs 2600 (*green*), 2500 (*cyan*) and 2818 (*peru*) are also pictured.

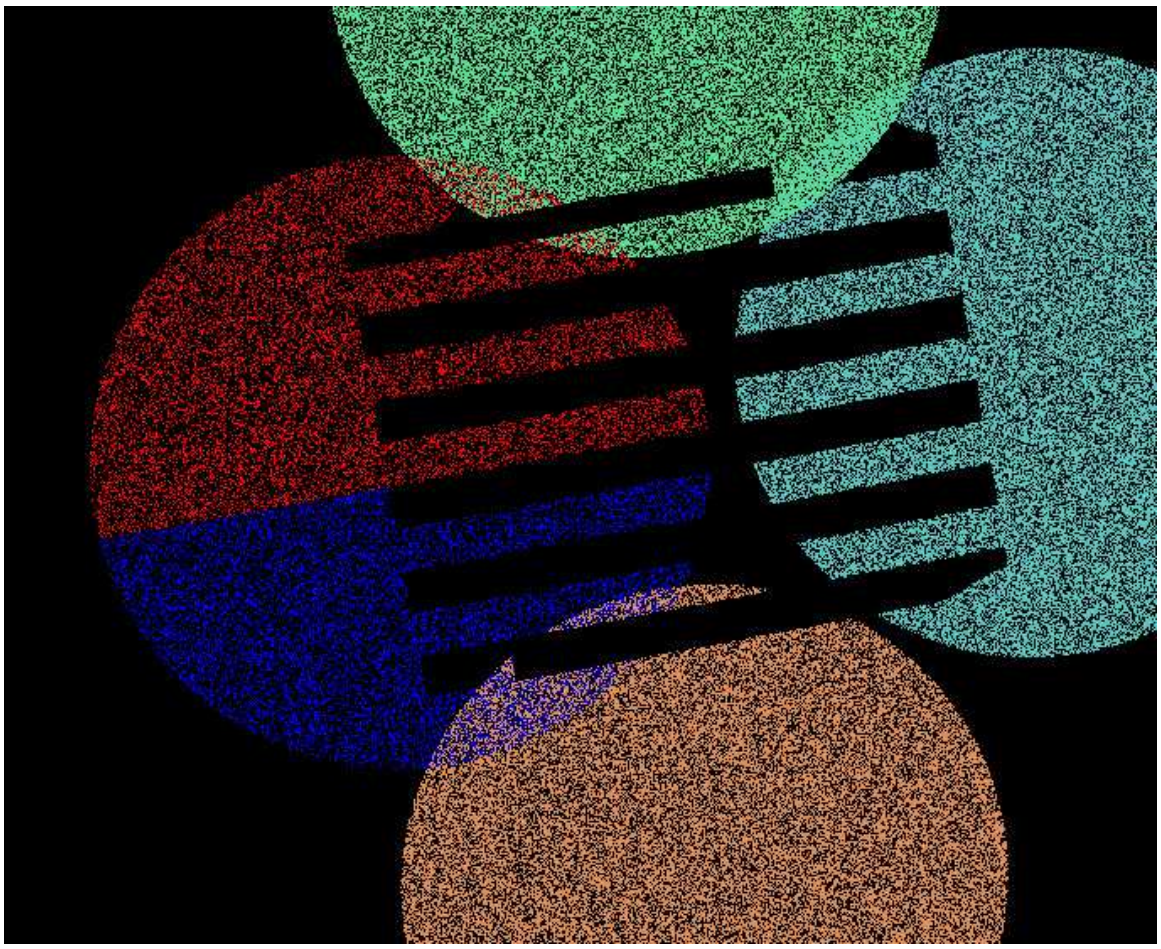

Figure 5.6: Illustration of a constraint condition that defines the upper boundary of one of Bad Area 1's SEGMENTs. Points that lie within the union of DR7 TILE 2499's SECTORs and which also satisfy the constraint condition are colored in blue. Were these points not confined to TILE 2499, they would fill an entire hemisphere. Clockwise from the top, the union of SECTORs within DR7 TILEs 2600, 2500, 2818 and 2499 are also pictured.

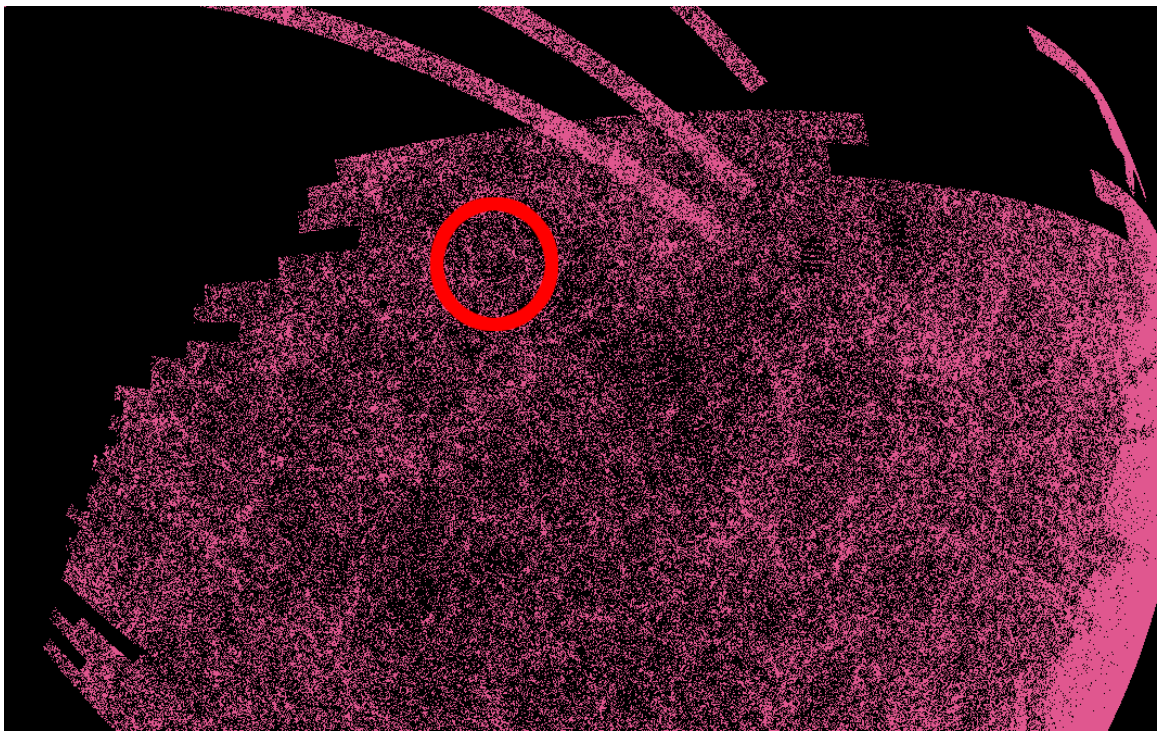

Figure 5.7: The location of Bad Area 2 is circled in red. The approximate position of the region is $[RA, dec] \approx [152°, 58°]$. MGS targets are represented by magenta pixels.

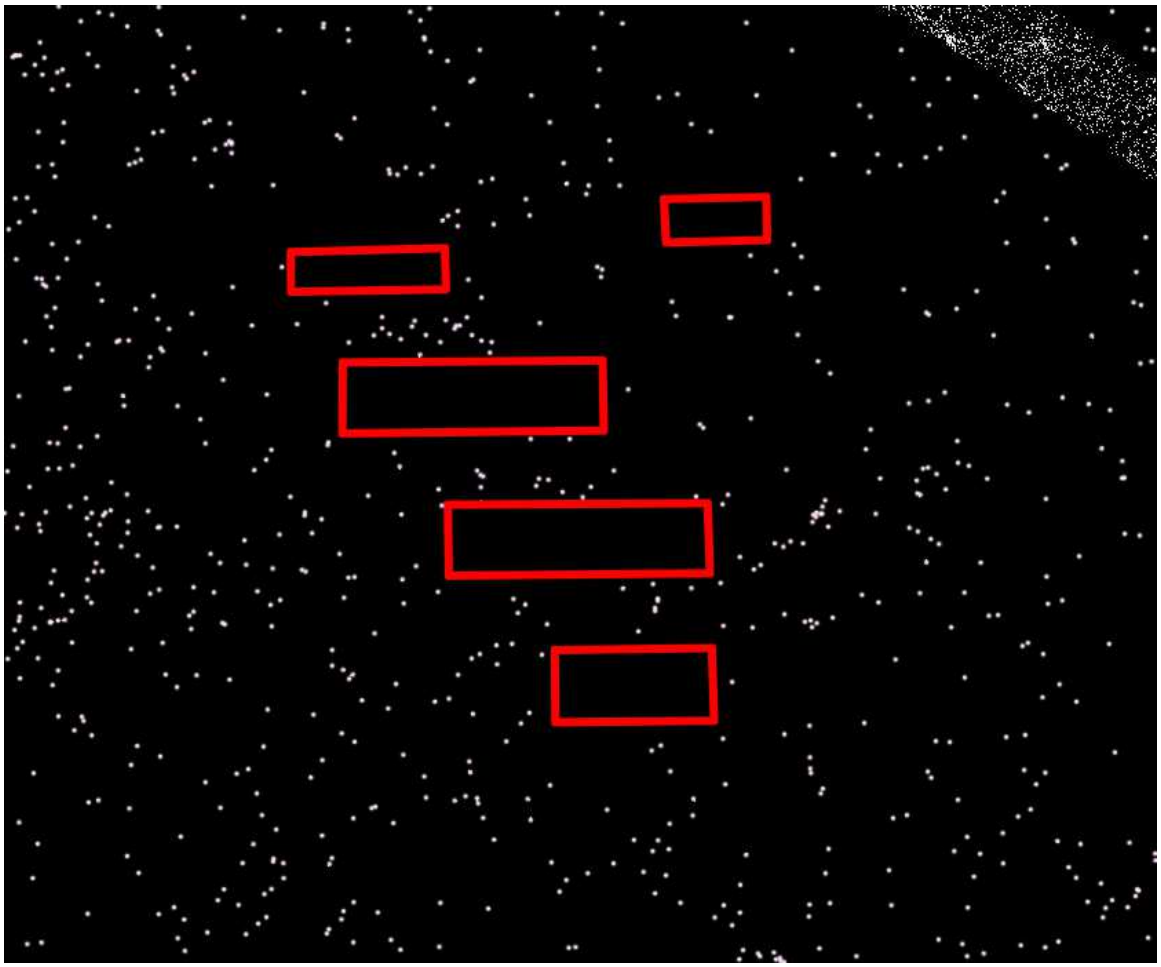

Figure 5.8: Close-up view of Bad Area 2. The five PRIMARY SEGMENT portions that require removal from the photometric footprint are boxed in red. MGS targets are represented by white pixels.

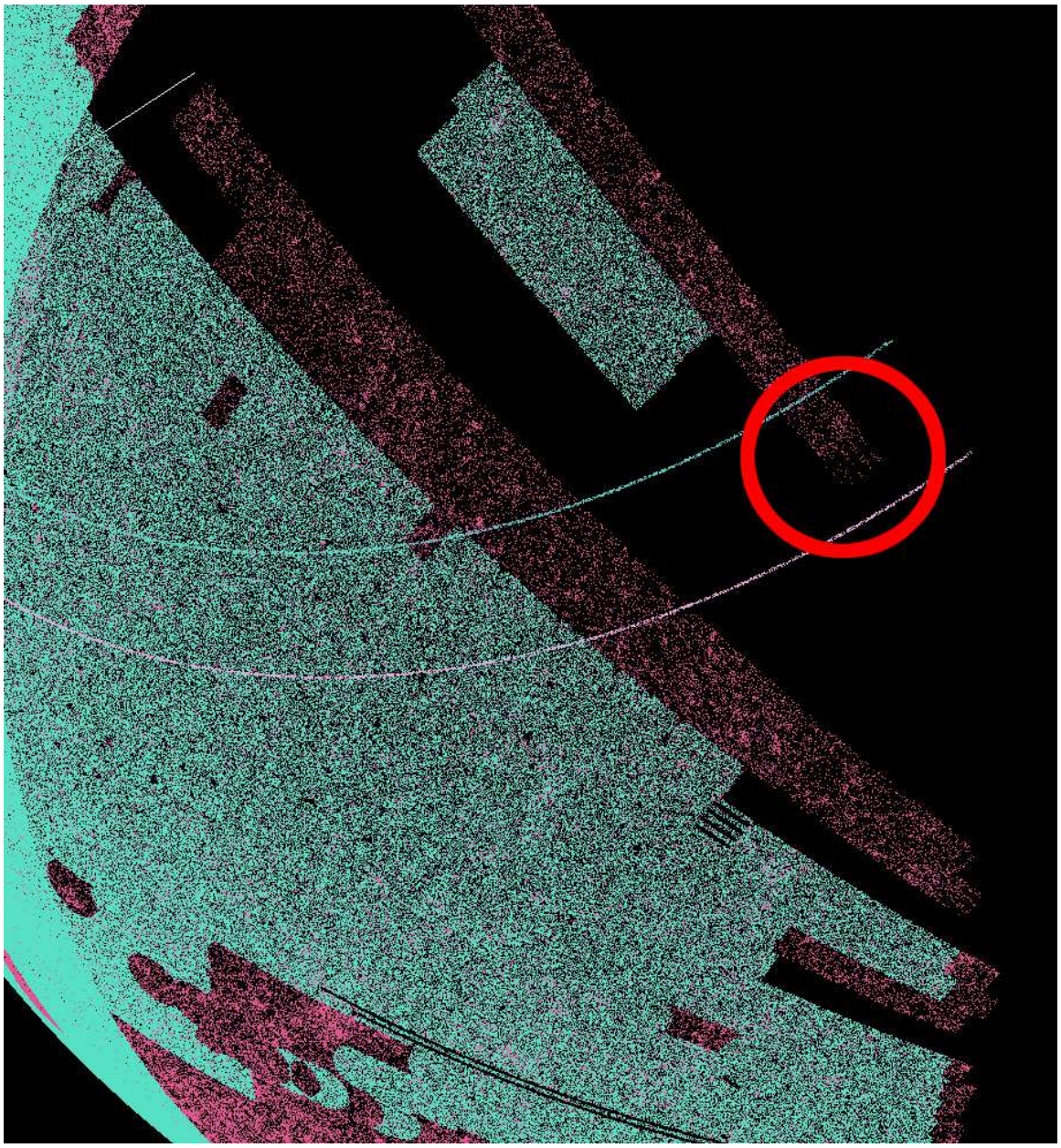

Figure 5.9: The location of Bad Area 3 is circled in red. The region's position is $[RA, dec] \approx [269°, 47°]$ where the curved lines are declinations of $45°$ (*bottom*) and $55°$ (*top*). The DR6 improved spectroscopic footprint is colored in cyan. MGS targets, which mark the extent of the photometric footprint, are colored in magenta.

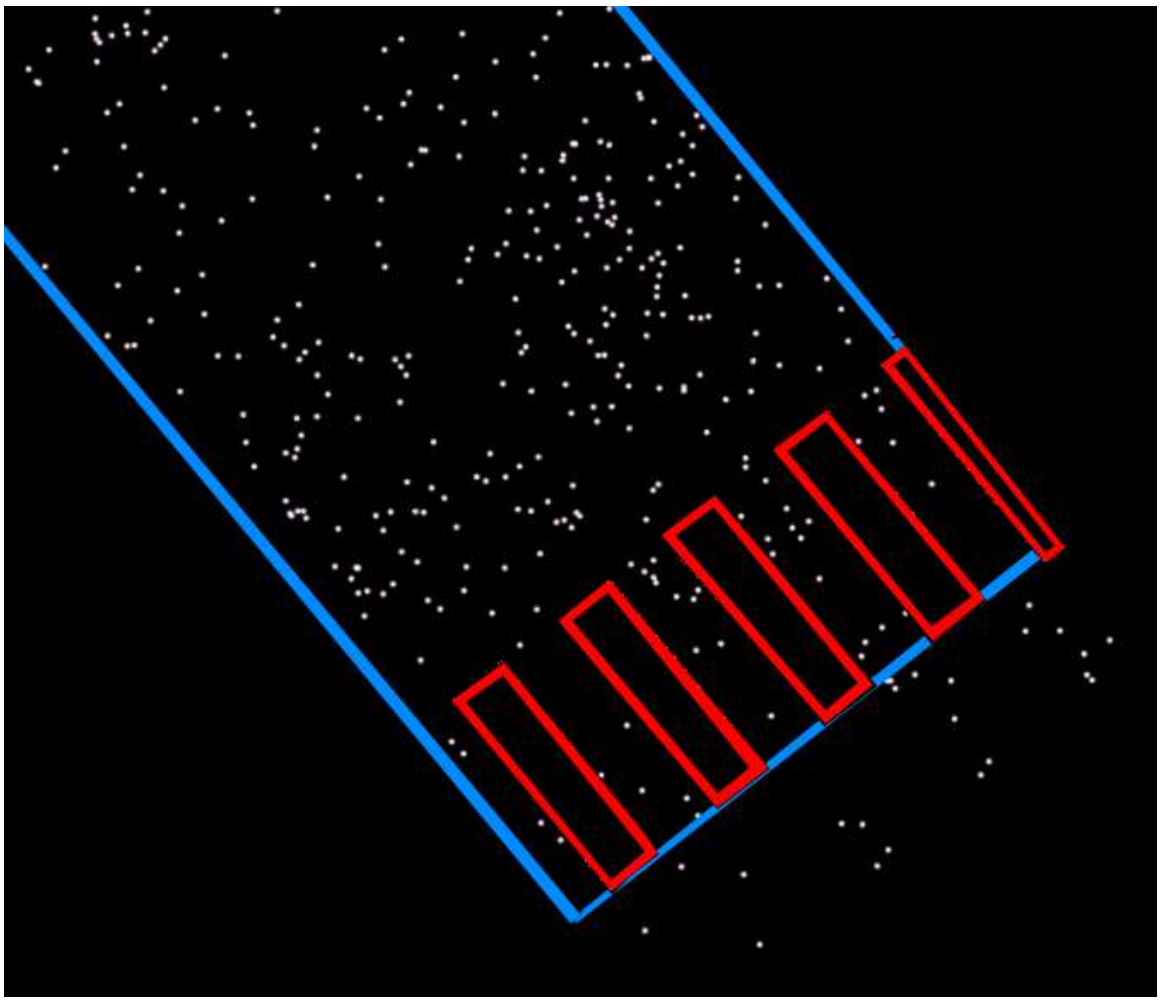

Figure 5.10: Close-up view of Bad Area 3. PRIMARY SEGMENTs that require removal are boxed in red. These are portions of DR6 SEGMENTs 1766 (*lower left*) through 1770 (*upper right*). MGS targets, represented by white pixels, are conspicuously absent from these regions. The boundary of the PRIMARY SEGMENT footprint is outlined in blue. Galaxies outside this boundary are secondary targets and are not considered in our analysis.

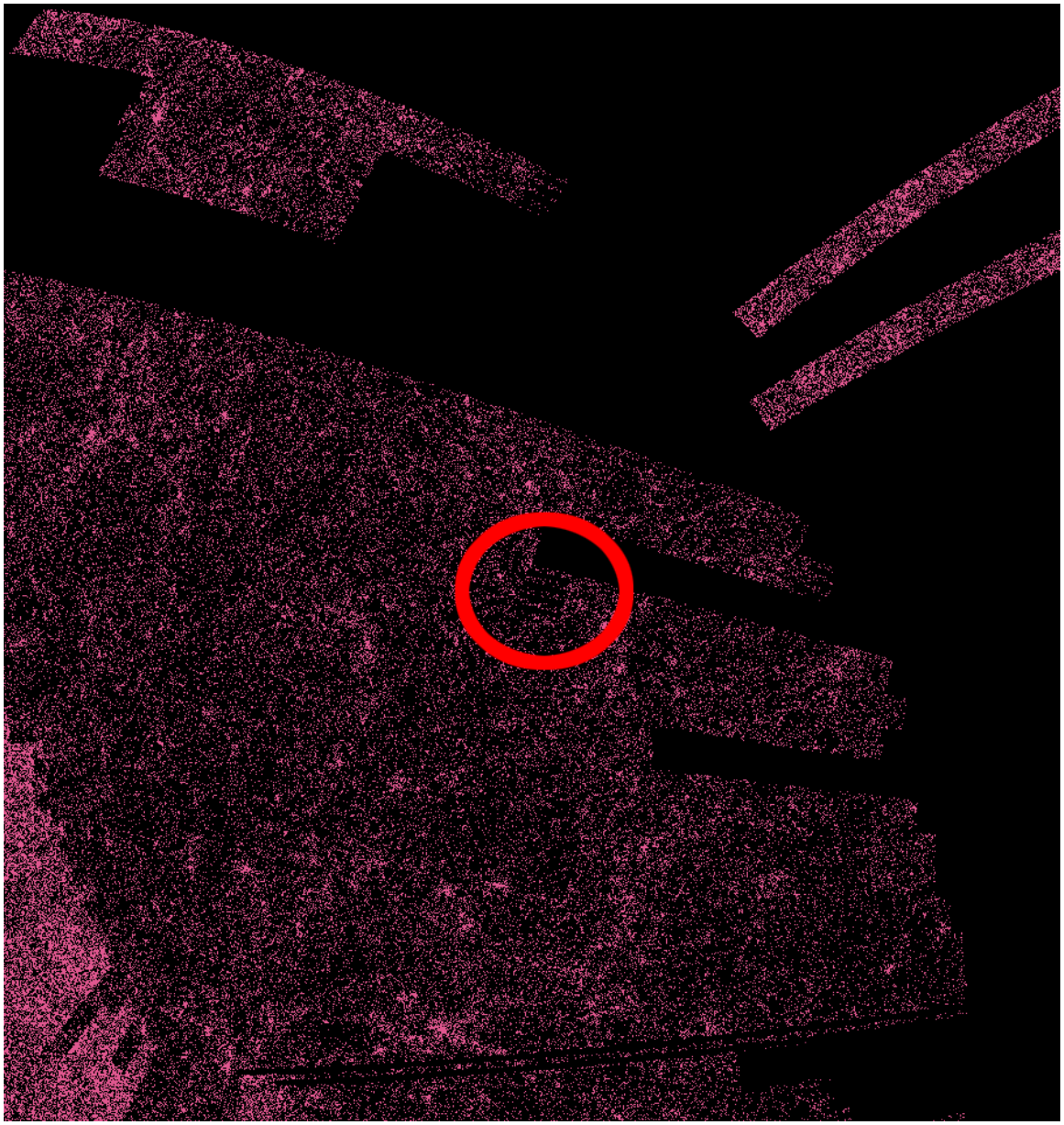

Figure 5.11: The location of Bad Area 4 is circled in red. The approximate position of the region is $[RA, dec] \approx [255°, 37°]$. MGS targets are colored in magenta.

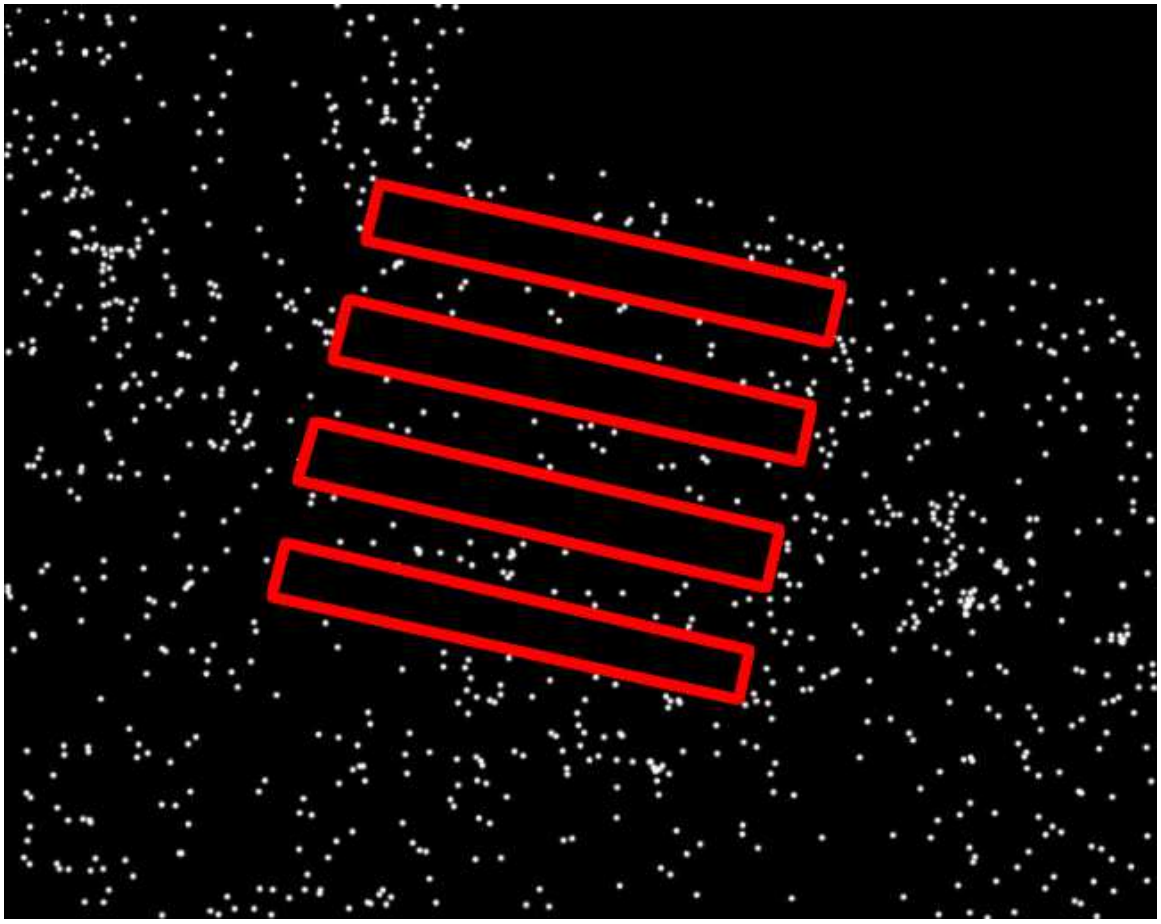

Figure 5.12: Close-up view of Bad Area 4. The four SEGMENT portions that require removal from the photometric footprint are boxed in red. MGS targets represented by white pixels.

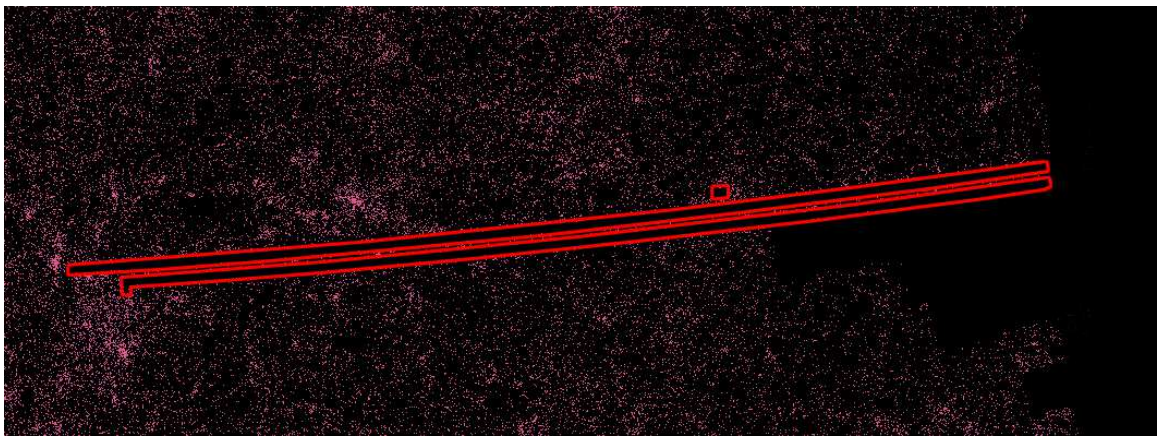

Figure 5.13: View of Bad Area 5. The union of the four rectangular regions that require removal from the photometric footprint are outlined in red. MGS targets are represented by magenta pixels.

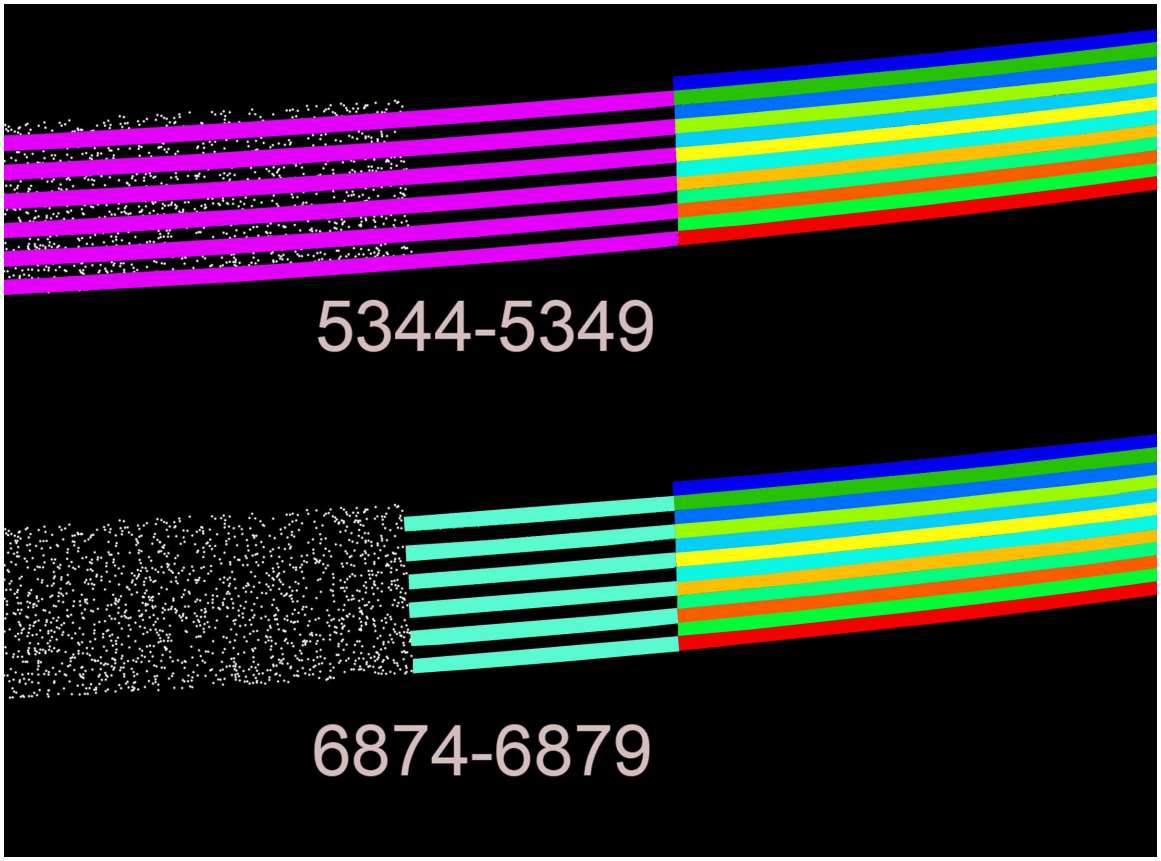

Figure 5.14: Visualization of the Bad STRIP. The top image shows the junction of the two highest declination STRIPEs in the southern hemisphere. The rainbow colored SEGMENTs on the right belong to their own PRIMARY. The six SEGMENTs colored in magenta represent SEGMENTs 5344 through 5349. Randomly placed white dots mark the boundary of the photometric footprint. The bottom image shows the same region of space except now the six SEGMENTs colored in cyan represent SEGMENTs 6874 through 6879.

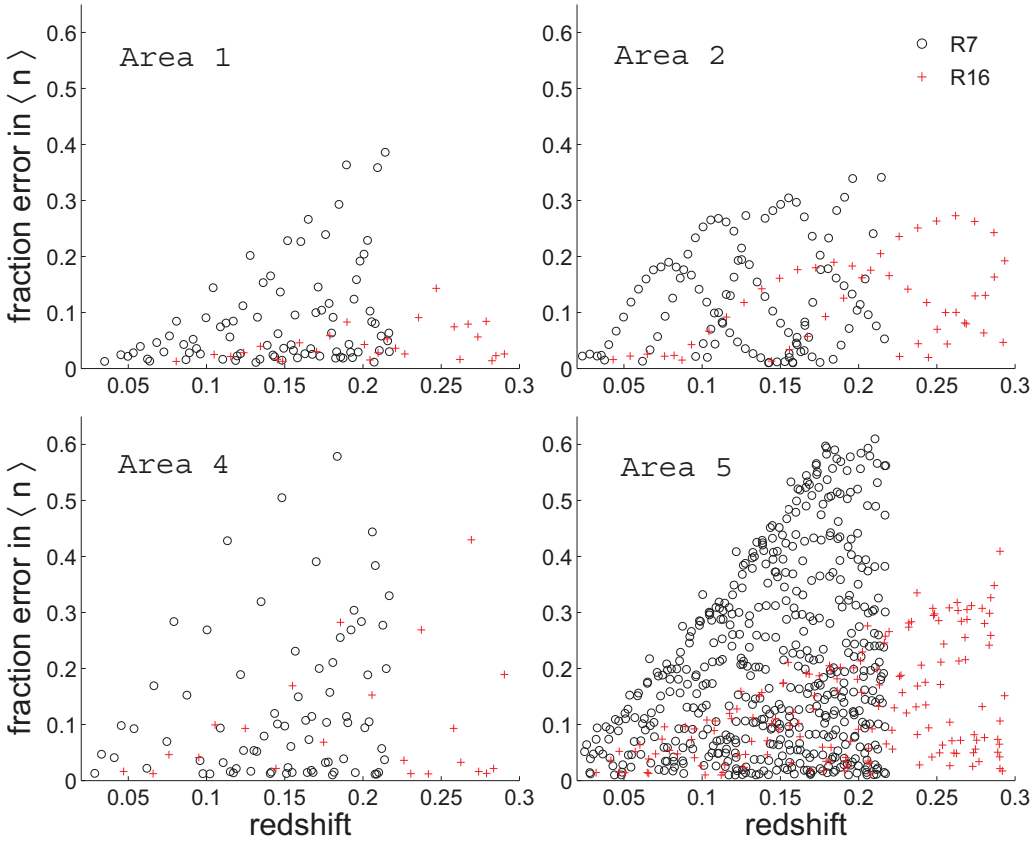

Figure 5.15: The errors $\left(\beta_{PS}^{(0)} - \beta_{PS}\right)/\beta_{PS}$ in the expected number of galaxies $\langle n \rangle$ within cells that intersect Bad Areas 1, 2, 4 and 5 as a function of redshift. Results for R7 and R16 cells are presented. The bell curve features in Bad Area 2 result from the regular geometry of the cells positioned by the HCP arrangement and do not reflect any sort of hidden feature.

5.2 Spectroscopic Footprint

The DR6 spectroscopic footprint is defined to be to the union of all DR6 SECTORs. However, there are hundreds of individual regions where this simple definition fails. This section addresses the challenge of repairing the spectroscopic footprint to the point where its completeness is not subject to gross undersampling biases or incorrectly placed boundaries.

We begin by identifying five regions that contain MGS galaxies (with spectra), but which lie outside the union of SECTORs. We provide constraint conditions to reintroduce these areas to the footprint. Next, we show how a massive, low declination CHUNK made it into the union of SECTORs even though no spectroscopically detected objects lie with it. Then we report on a portion of an undersampled SEGMENT that must be removed.

We continue by studying the hundreds of undersampled SECTOR-type regions within the DR6 footprint. We show how these regions are located, defined and removed. We conclude by visualizing the improved spectroscopic footprint and reporting the stark statistical differences between MGS targets inside its area and those trimmed from it.

5.3 Inclusion/Exclusion Regions

Figures 5.16 and 5.17 visualize two of the lowest declination areas within the northern hemisphere. Figure 5.16 contains three regions labeled A, B and C. Each of these regions lies outside the union of SECTORs but all contain MGS galaxies. Figure 5.17 visualizes

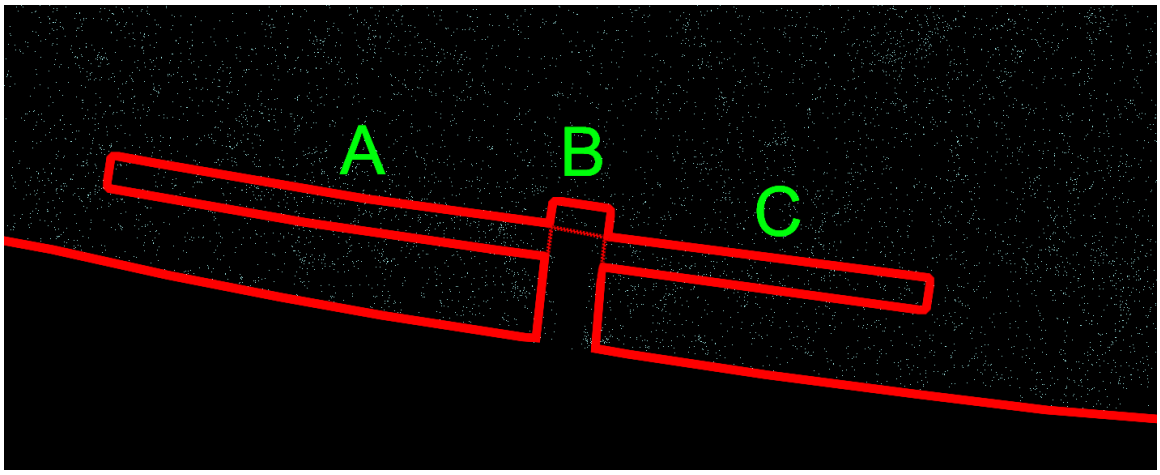

Figure 5.16: Visualizations of regions A, B and C that are added to the improved spectroscopic footprint. The red line marks the boundary of the union of DR6 SECTORs. Regions A, B and C share this boundary on three sides and the dotted red line boundaries on their fourth. MGS galaxies are marked by blue pixels. About 60 of these galaxies exist outside the footprint below region B. However, no constraint conditions could be found in the database to mark the boundaries within which they are contained. These galaxies are therefore excluded from the improved spectroscopic footprint.

a similar situation whereby areas D and E have spectroscopically detected objects within them, yet somehow lie outside the spectroscopic footprint. Table 5.2 reports the four constraint conditions needed to reintroduce each to the improved spectroscopic footprint.

In Figure 5.18, MGS galaxies are superimposed on top of the spectroscopic footprint. Areas visible in cyan lie within the union of SECTORs yet contain no MGS galaxies. The smallest of these areas in the interior of the survey largely lie in the compliment of TILE intersections and are an expected byproduct of the SDSS survey strategy.

Larger cyan areas are indicative of regions that are undersampled. The most prominent of these is the large rectangular area at the bottom of the image. This region is CHUNK 113. Despite lying in the union of DR6 SECTORs, it contains no MGS galaxies. As such,

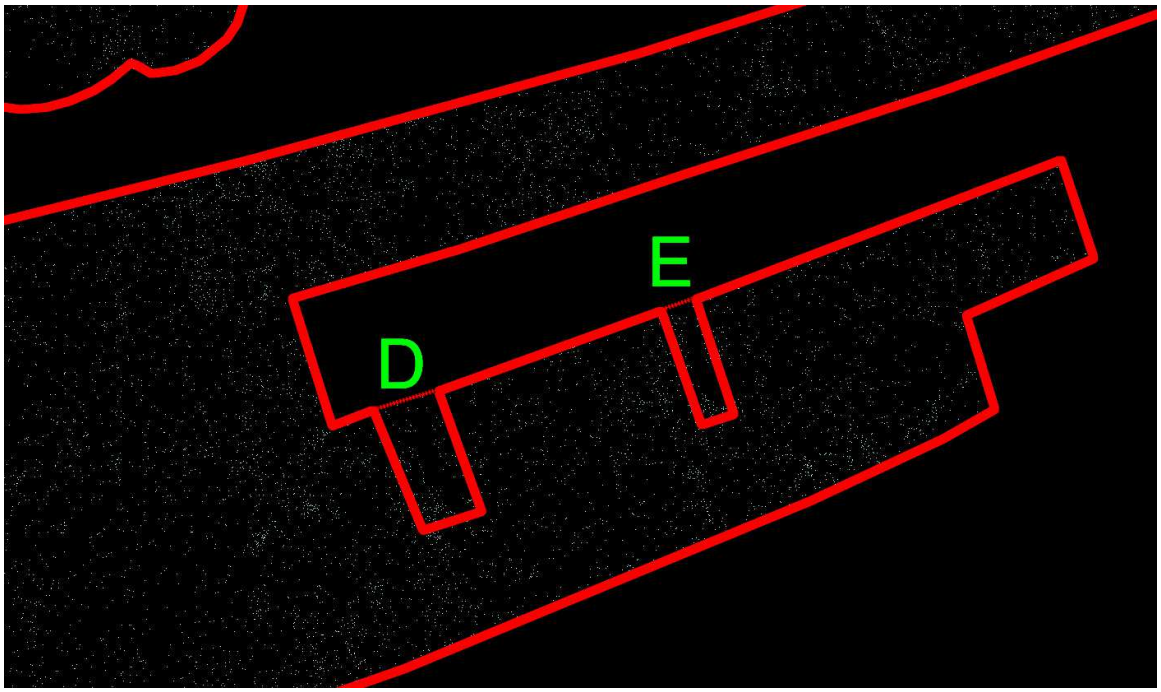

Figure 5.17: Visualization of regions D and E that are added to the improved spectroscopic footprint. Lines and pixels are the same as in Figure 5.16.

CHUNK 113 is deleted from the improved spectroscopic footprint.

The lowest R.A. section of DR6 SEGMENT 5417, pictured in Figure 5.19, is also removed because its area to the left of the DR7 TILE pictured contains no MGS galaxies. This region is defined by the intersection of SEGMENT 5417 with any of the following DR6 SECTORS: 39201, 39205 or 39212.

Finally, the portion of the STRIPE spanned by SEGMENTs 6874-6879 in Figure 5.14 is removed. Because the spectroscopic footprint is a subset of the photometric footprint, areas removed from the latter must also be removed from the former. The ambiguous, overlapping region definitions that prompted this removal were covered in §5.1.1.

Area	Constraint 1	Constraint 2	Constraint 3	Constraint 4
A	-33111	-33094	33101	5264
B	269	33111	5252	5251
C	-33111	-33094	5265	33092
D	269	270	5255	5256
E	269	270	5259	5260

Table 5.2: DR6 constraint conditions that define regions A through E. A negative sign indicates that the $[x, y, z, c]$ coordinates had their signs reversed to ensure the constraint was on the correct side of the halfspace. Points that satisfy all four of an area's constraints lie within that rectangular region.

5.3.1 Undersampled SECTORs

The spectroscopic footprint can be separated into two kinds of areas. The first, which lies mostly in the interior of the survey, was observed in such a way that the percentage of MGS targets lacking spectra is approximately 6% for fiber collided galaxies (Strauss et al., 2002) and 20% overall. The second, which lies mostly on the edges of the survey, is comprised of regions where further spectroscopic measurements were planned for DR7. Here, the percentage of targets lacking spectra routinely ranged between 50-100%. These two types of areas have vastly different statistical properties and deserve specialized handling.

An example of the second type of area is shown in Figure 5.20. The portion of the survey pictured lies on the edge of the DR6 spectroscopic footprint. Most of the visible gaps in spectroscopic coverage were eventually filled in during DR7. Areas shaded in gray lie within the union of DR6 SECTORs yet contain no MGS galaxies.

The shapes of these areas appear to be formed from the intersections of circular TILEs. Upon overlaying TILEs placed during DR7, we find this is indeed the case. The TILE in

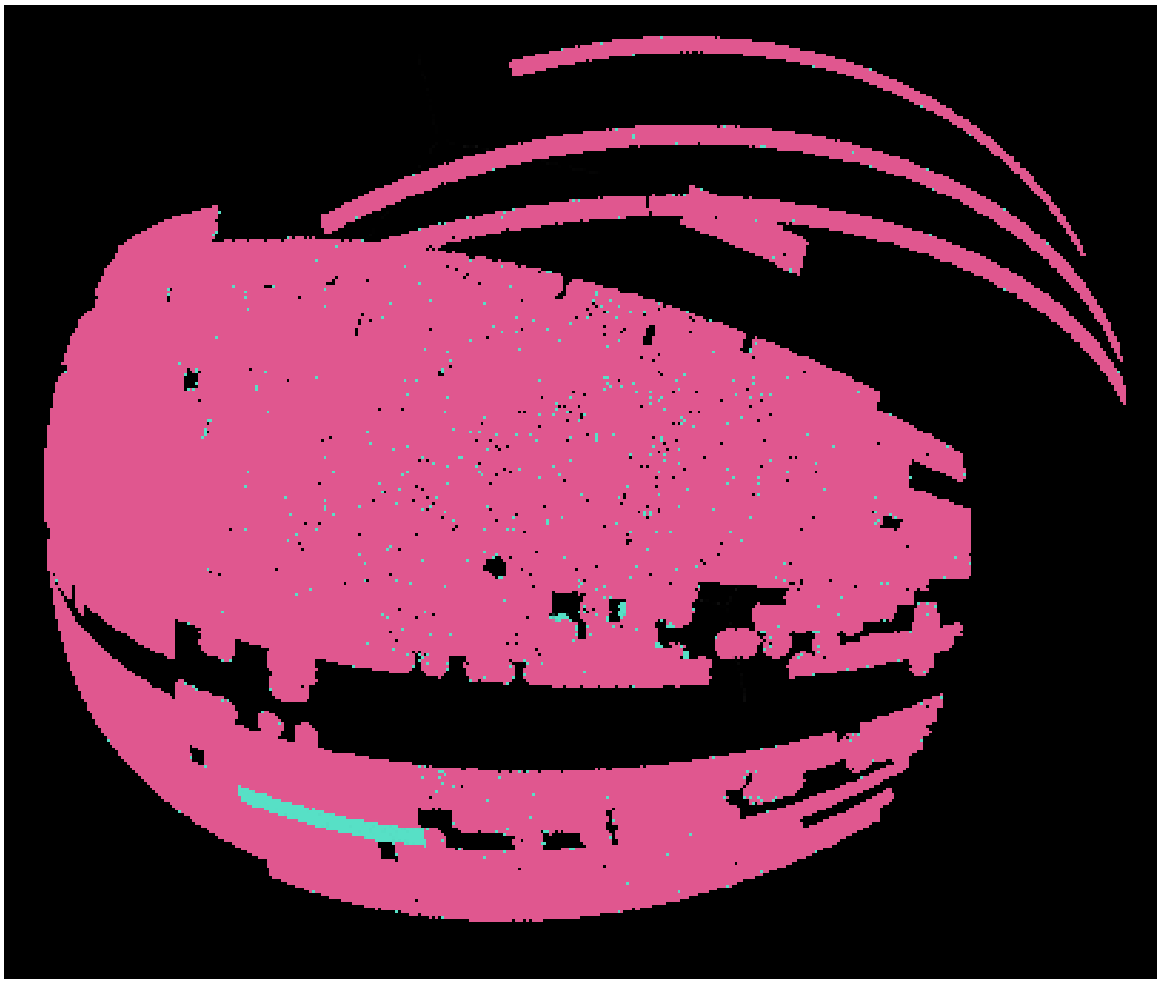

Figure 5.18: Comparison between the DR6 spectroscopic footprint (*cyan*) and MGS galaxies (*magenta*). The large rectangular area in the lower declination region of the northern hemisphere is the area covered by CHUNK 113.

Figure 5.20 contains five DR7 SECTORs that clearly overlap the undersampled region. We emphasize that this region cannot be defined through the geometric descriptions provided in the DR6. This indicates that research published using DR6 data prior to the release of DR7 would almost certainly have been unable to properly account for these regions.

Through a tedious process of visually comparing the positions MGS galaxies against the extent of the spectroscopic footprint (as provided by Monte Carlo points filtered through

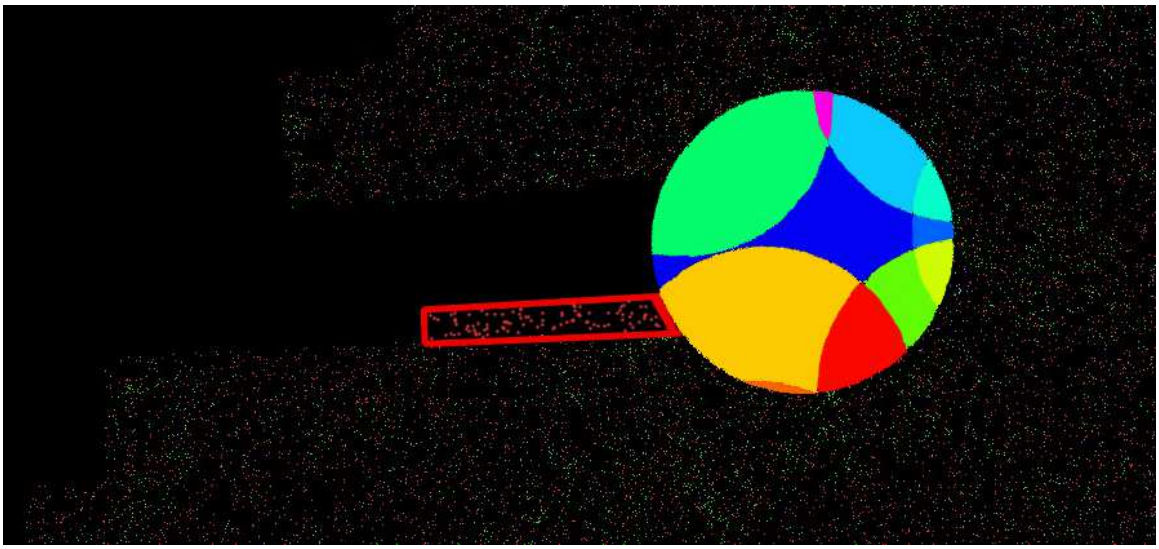

Figure 5.19: An area at the edge of DR6 SEGMENT 5417 that is removed from the improved spectroscopic footprint is bounded in red. This area is defined to lie within union of DR6 SECTORs (marked empirically with red pixels) but contains no MGS galaxies (marked in green pixels). The DR7 TILE that defines this area's right edge is displayed with its SECTORs individually colored.

DR6 SECTORs) we were able to identify 183 areas covered by DR7 SECTORs that appeared to be either not sampled or significantly undersampled during DR6. The removal of these regions resulted in what we refer to as the *better spectroscopic footprint*.

The selection process operated according to a few principles. First, because undersampled regions were located exclusively near the edges of the DR6 spectroscopic survey, the majority of our attention was focused here. This limited view helped differentiate between regions that were undersampled and those that were merely underdense.

Second, we overlaid suspect areas with DR7 TILEs and SECTORs. If the shapes of these areas visually matched the shapes of the DR7 geometry, the latter's constraint conditions were gathered so those SECTORs could be removed.

Third, SECTORs that lay within contiguous "gray areas" as in Figure 5.20 were re-

moved as a group. It is possible that some of the smaller SECTORs within that group were adequately sampled, but contained a paucity of galaxies through cosmic variance. We give such SECTORs no benefit of the doubt and assume that those in the gray areas were similarly treated by the tiling algorithm.

Fourth, there were borderline cases in which it was unclear whether an area was undersampled or just underdense. Under these circumstances we defaulted towards removing SECTORs because 1) the primary purpose of improving the footprint is to ensure that cells are only placed in positions that we know for sure have good spectroscopic coverage and 2) measures of galaxy surface density are less sensitive to the exclusion of good SECTORs than to the inclusion of bad ones. The extent of MGS galaxies is vast, and while a handful may be unnecessarily excluded, it comes with the peace of mind that the remaining spectroscopic footprint has relatively uniform sampling properties and can be treated similarly.

A view of the edge of the better spectroscopic footprint is presented in Figure 5.21. This picture of MGS objects (as opposed to the MGS galaxies that were used previously) reveals more regions of high incompleteness. The anisotropies discovered in this manner led to a second round of corrections that eliminated an additional 120 SECTORs.

In total, the areas covered by 303 DR7 SECTORs were removed to produce the *improved spectroscopic footprint*. We have made a list of these SECTORs available at this link.

5.3.2 Improved Spectroscopic Footprint

In summary, the final DR6 *improved spectroscopic footprint* is created through these steps:

- Include the union of all DR6 SECTORs
- Include the areas in regions A, B, C, D and E
- Remove CHUNK 113
- Remove the area within the intersection DR6 SEGMENT 5417 and DR6 SECTOR 39201 or 39205 or 39212 (all yield the same result)
- Remove the STRIPE containing DR6 SEGMENTs 6874-6879
- Remove the area in all 303 undersampled DR7 SECTORs

Figure 5.22 offers a view of the newly trimmed footprint. Taken together, the improvements to the spectroscopic footprint reduce its area from 6860 $\deg^2$ (Adelman-McCarthy et al., 2008) to 6621.3$\pm$1.8 $\deg^2$. This constitutes an adjustment of about 239 $\deg^2$, or 3.6%, from the commonly reported value.

Finally, we note that our improved spectroscopic footprint is almost certainly imperfect. As seen in Figure 5.23, small border regions containing no-redshift MGS objects still appear relatively overdense. The same holds for a low R.A., high declination CHUNK in the northern hemisphere. However, anisotropies in these regions are relatively minor and do not warrant the removal of a large chunk of the northern sky.

	Number of Targets		**Number per deg^2**		
	Union of DR6 SECTORs	Improved Spectroscopic Footprint	Union of DR6 SECTORs	Improved Spectroscopic Footprint	Discarded Spectroscopic Footprint
Pristine Galaxies	$479,419$	$474,651$	69.886	71.69 ± 0.02	19.975 ± 0.005
No-Redshift Objects	$114,038$	$99,964$	16.624	15.097 ± 0.004	58.961 ± 0.016
Low-Quality Objects	$22,794$	$22,517$	3.3227	3.4007 ± 0.0009	0.9510 ± 0.0003
Spectroscopic Completeness			0.778	0.795	0.250

Table 5.3: Comparison between galaxy counts, densities and spectroscopic completenesses inside and outside the improved spectroscopic footprint.

5.4 Spatial Distribution of MGS Targets

The trimming of SECTORs required to generate the improved spectroscopic footprint affects both the number counts and densities of MGS targets within its boundary. Table 5.3 summarizes these changes while Figure 5.23 displays the angular distribution of the three types of MGS targets.

There are substantial statistical differences between MGS targets inside and outside the improved spectroscopic footprint. While spectroscopic completeness anisotropies are to be expected, the magnitudes of the disparities reported in Table 5.3 validate the decision to separate the two areas.

In short, these results suggest our footprint trimming has the desired effect. An improved area has been created in which the density of targets with spectra has increased while the density of targets without spectra has decreased. Overall, a 3.6% reduction in the

spectroscopic footprint area led a 2.2% increase in the spectroscopic completeness of the survey.

Furthermore, the number density of MGS objects is about 4 times greater outside the improved spectroscopic footprint than inside. In comparison, the density of MGS galaxies is about 3.6 times smaller outside than inside.

Another benefit is that, for a given minimum β_{SPEC}, the absolute number of MGS objects within the cells' projections has decreased. Trimming the footprint removed almost 3 times more MGS objects than MGS galaxies even though the latter are 4 times as prevalent overall. This improvement reduces the uncertainty in galaxy counts per cell, particularly for those cells on the survey's edge, without eliminating an undesirably high amount of useful redshift information.

Finally, as we will argue in Chapter 6, there exist methods for estimating the radial positions of MGS objects (e.g. nearest neighbor, two-point correlation function smearing) that depend upon the depth of the nearest angular neighbor. Failure to improve the spectroscopic footprint weakens those methods and adds unnecessary uncertainty to overdensity measurements on the survey's edges.

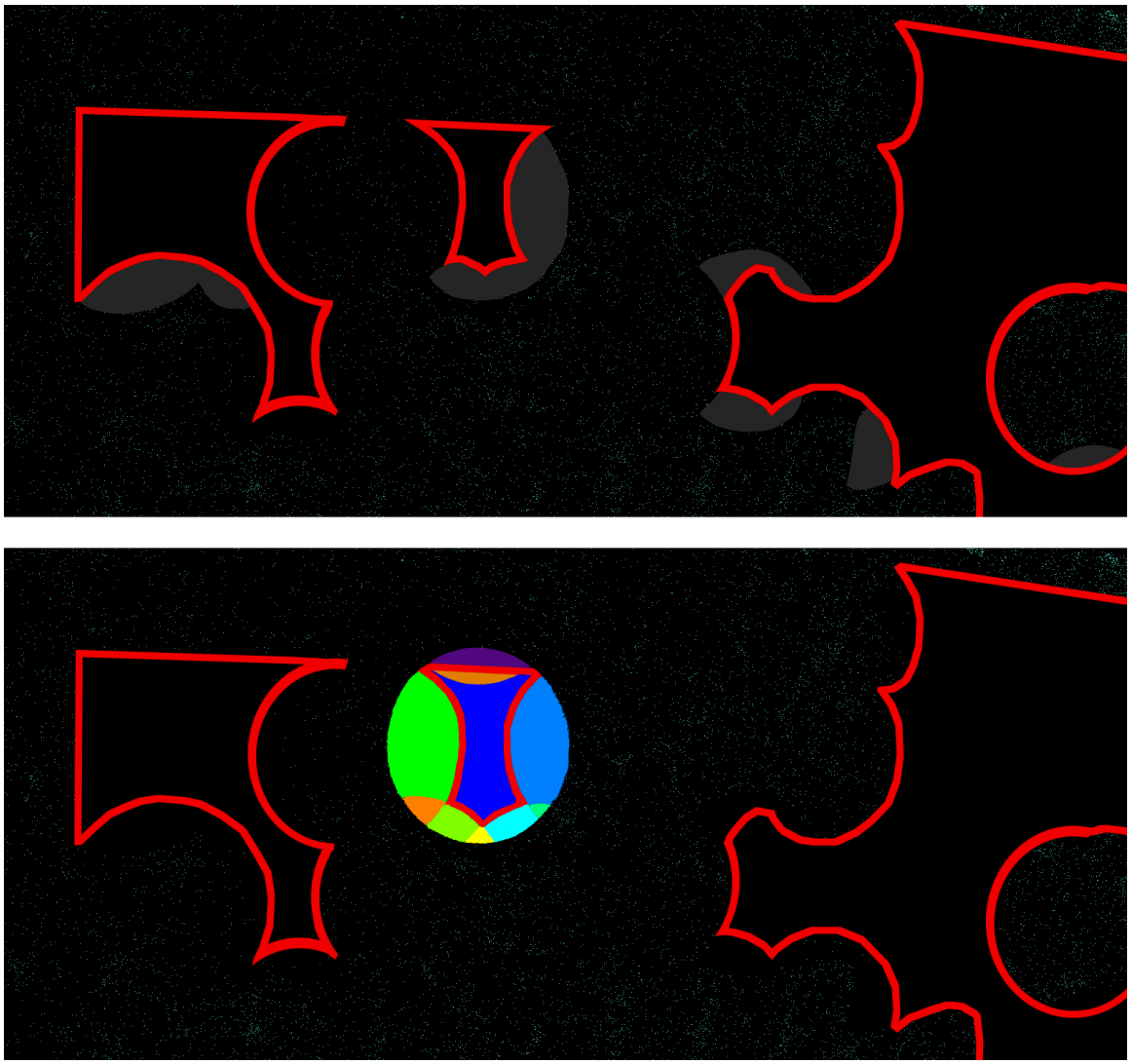

Figure 5.20: Illustration of undersampled SECTORs in the DR6 spectroscopic footprint. The red lines mark the boundary of the union of DR6 SECTORs while the blue pixels indicate the positions of MGS galaxies. In the top panel, select regions within the spectroscopic footprint that contain no MGS galaxies are shaded in gray. In the bottom panel a DR7 TILE is superimposed. Five of its SECTORs overlap the undersampled area in the spectroscopic footprint. The region pictured lies approximately in the range $RA \in [205°, 220°]$ and $dec \in [25°, 35°]$.

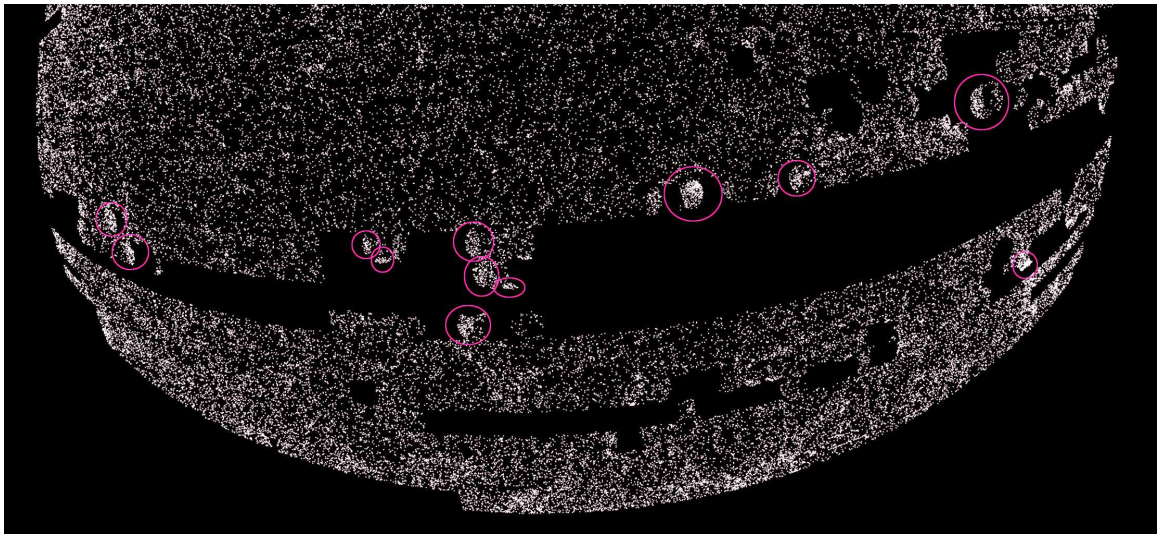

Figure 5.21: Distribution of MGS objects in the better spectroscopic footprint. Areas where the number densities of objects appear to be significantly higher than average are circled in pink. The DR7 SECTORs that cover these areas are ultimately removed from the spectroscopic footprint.

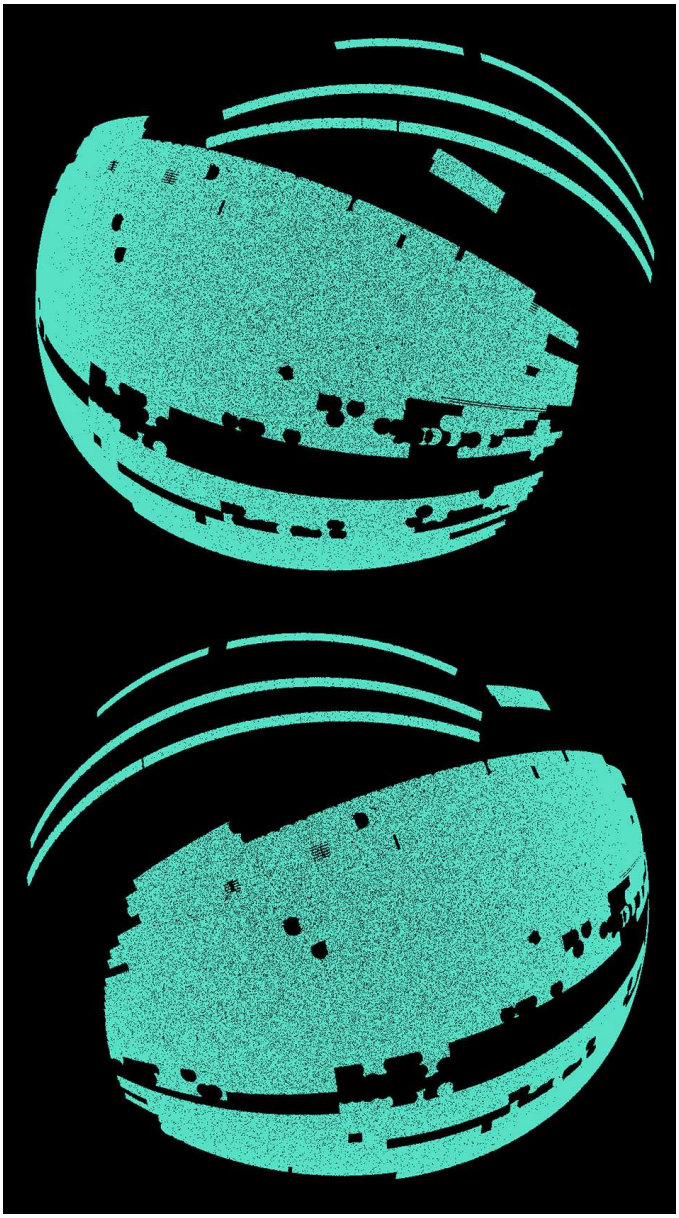

Figure 5.22: Two perspectives of the DR6 improved spectroscopic footprint as projected onto the celestial sphere. The footprint is visualized empirically using full sky angular randoms filtered through DR6 SECTORs, followed by the corrections described in this chapter. The appearance of many of the tiny holes in the survey interior is a result of the limited resolution of the angular randoms and do not necessarily represent actual holes in the footprint. Regions near the edges of the footprint have been trimmed so that the remaining areas have approximately the same angular completeness.

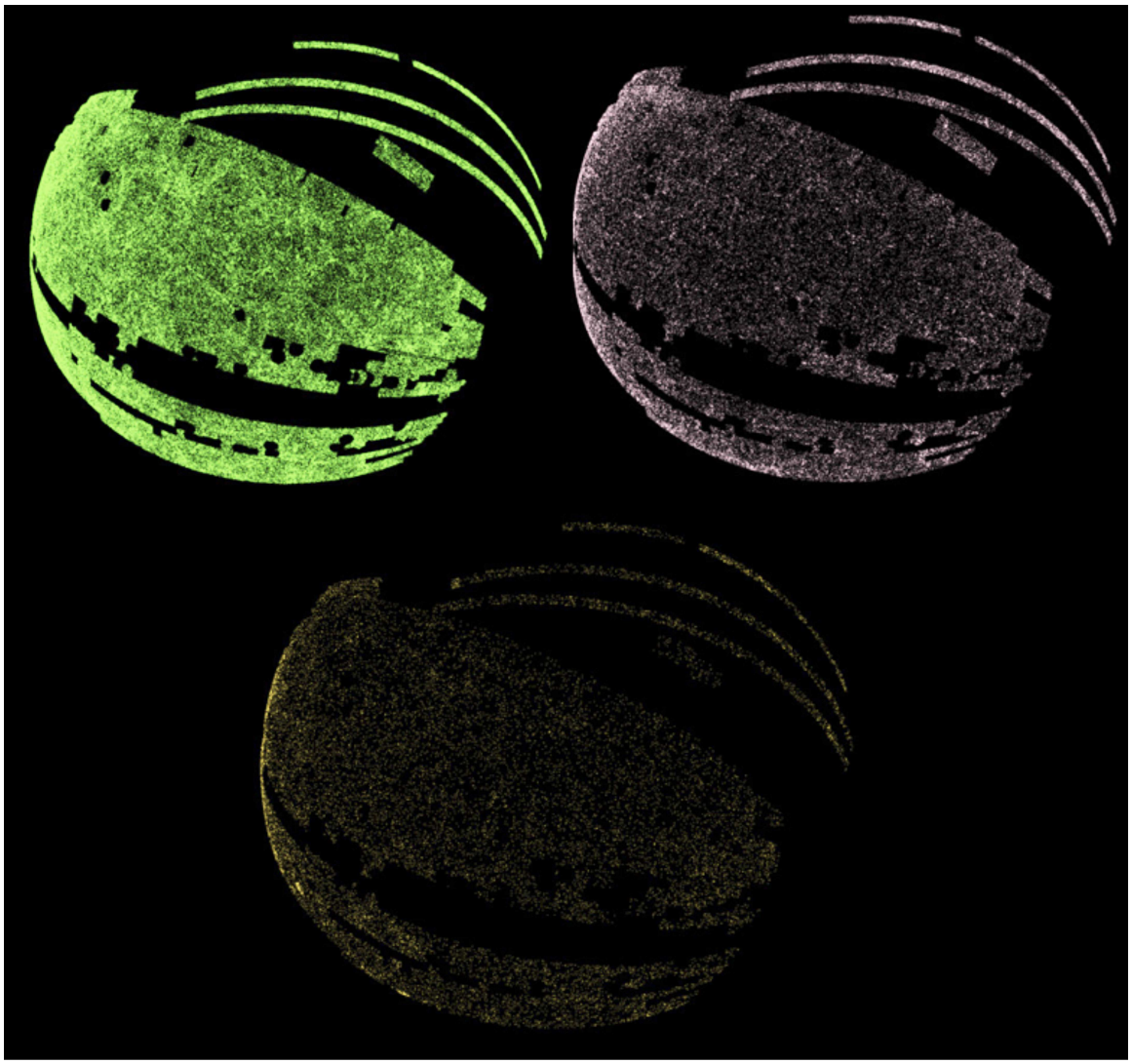

Figure 5.23: Angular distribution of MGS targets within the improved spectroscopic footprint. Pictured are the MGS galaxies (*upper left*), MGS objects without spectra (*upper right*) and MGS objects with low-quality spectra (*bottom*).

Chapter 6

Counting Galaxies in Cells

Measuring the large scale distribution of galaxies is impossible without quantifying their radial depths. No observable is more strongly correlated with a galaxy's distance than its spectroscopic redshift. Yet as we have shown, approximately 20% of MGS targets within the spectroscopic footprint are without spectra. The situation is even worse where galaxy density is high. Yoon et al. (2008) cited a 30%–40% incompleteness rate in dense regions while searching for galaxy clusters in three dimensions. In their study of galaxy clusters, von der Linden et al. (2007) discovered that the central galaxy in 30% of clusters was missing a redshift.

Recall that MGS targets can go spectroscopically unobserved for a number of reasons. An engineering limitation prohibits spectroscopic fibers from being placed arbitrarily close on a tile, leading to fiber collisions. The tiling algorithm may undersample a region in one data release with the expectation that the difference will be made up in the next. A region

may contain so many targets that one or more tiles cannot accommodate enough fibers to detect them all. We refer the reader to Sections 2.1 and 2.2 for more details.

Spectroscopic incompleteness adds uncertainty to the number of galaxies contained within a volume and, by extension, the overdensity of that volume. The spatial distribution of overdensities determines the power spectra — linear, logarithmic, Gaussianized, and otherwise. As discussed in §3.2, power spectra are also a valuable tool in constraining a wide variety of cosmological parameters.

In this chapter, we study nine techniques that can be used to count galaxies in cells when a subset lacks spectroscopic redshifts. We include a discussion of the merits of each technique before applying them to MGS mock surveys. By the conclusion of the chapter, we will demonstrate the optimal strategy for counting targets depends on the region in which the objects are located, and the size and redshifts of the cells in which they are counted.

6.1 Counting Techniques

With this section, we present nine distinct counting techniques to account for the presence of MGS objects. To test these techniques, we simulate the distribution of MGS targets by randomly splitting the true MGS galaxy sample into two sets — *objects* that are stripped of their redshift information and *galaxies* that are not.[1] We then use the counting techniques

[1]To distinguish between MGS galaxies, MGS objects and the simulated *galaxies* and *objects* used to test the counting techniques, the latter set will be italicized for clarity.

to approximate the number count and overdensity in each cell. These measures are compared against the true counts, and conclusions are drawn regarding which techniques are most effective at a given redshift.

The nine counting techniques fall into three categories. The first is "discrete counting", in which every object is assigned a singular redshift. The second is "scaling", in which the number count of galaxies is scaled up by a factor related to either: A) a cell's spectroscopic completeness, or B) the volumes the cell occupies in the photometric and spectroscopic footprints. The third is "probabilistic smearing", in which an object's redshift is interpreted as a probability distribution function. The PDF is subsequently used to assign partial galaxy counts to cells along the object's line-of-sight. These distributions can be given by the selection function, two-point correlation function (2PCF), or photometric redshifts.

We classify objects by the environment they occupy. Objects that lie within the spectroscopic footprint, but lack spectra (perhaps due to fiber collisions or insufficient TILE depth) will be referred to as *interspersed objects*. Objects that lie outside the spectroscopic footprint, but inside the boundaries of cells will be known as *dark objects*. Finally, any object that lies in a large, contiguous area at a substantial distance from the spectroscopic footprint is referred to as an *external object*. The areas these objects occupy will be called *interspersed regions*, *dark regions*, and *external regions*, respectively. An example of a dark region is offered in Figures 6.1 and 6.2.

The list of counting methods that follows is by no means exhaustive. Others have used the color-magnitude relation (e.g Baum, 1959; Visvanathan & Sandage, 1977; Hogg et al.,

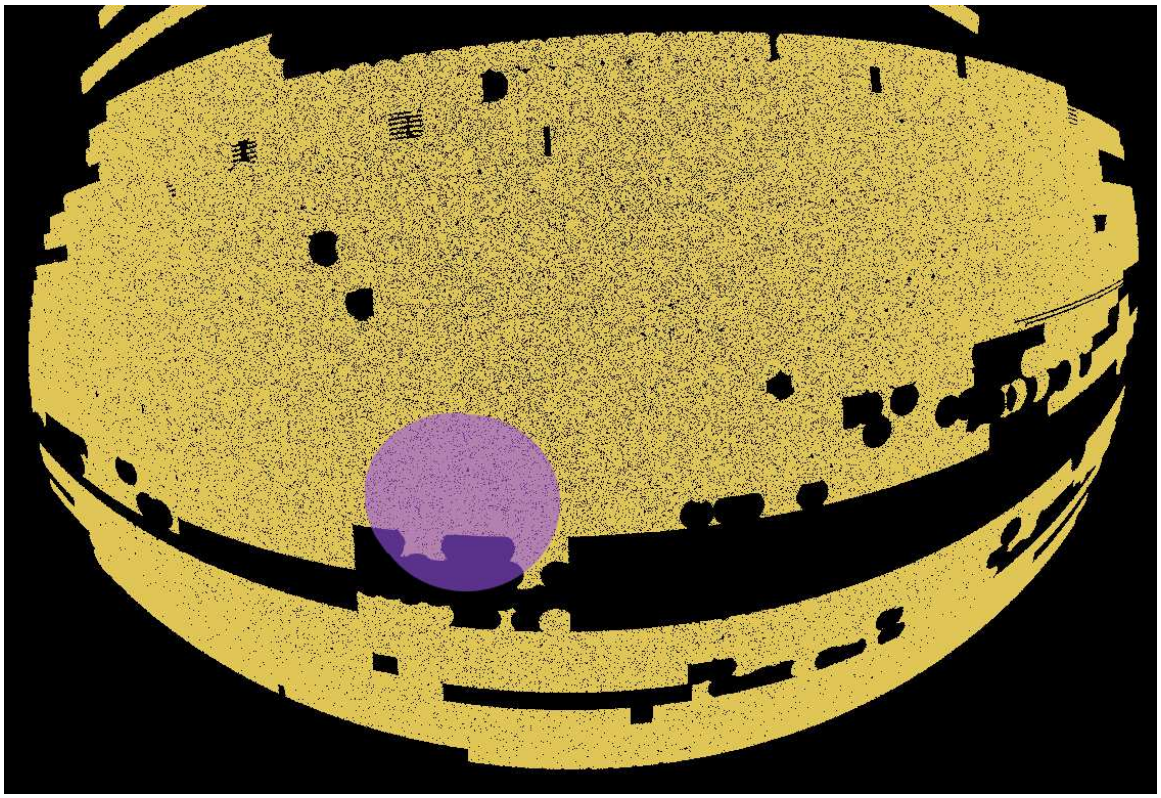

Figure 6.1: The circular projection of an R16 cell at $z = 0.037$ (*purple*) is superimposed atop a Monte Carlo visualization of the DR6 improved spectroscopic footprint (*yellow*). The portion of the cell that overlaps the black area, i.e. that which lies outside the spectroscopic footprint, is referred to as a *dark region*. For this cell, $\beta_{spec} = 0.8055$ and $\beta_{PS} = 1$. The fraction of the circular projection outside the spectroscopic footprint is 0.22.

2004; López-Cruz et al., 2004) to identify which angularly proximal early-type galaxies likely belonged to a particular cluster. Cunha et al. (2009) use a spectroscopic subsample of galaxies to assign an individual redshift probability distribution to each galaxy based on its photometry.

There is also no shortage of photo-z codes one can employ to correlate photometric properties with redshift. For instance, Hildebrandt et al. (2010) compare the performance of 19 such codes over 18 optical and near-infrared bands. Dahlen et al. (2013) conduct a

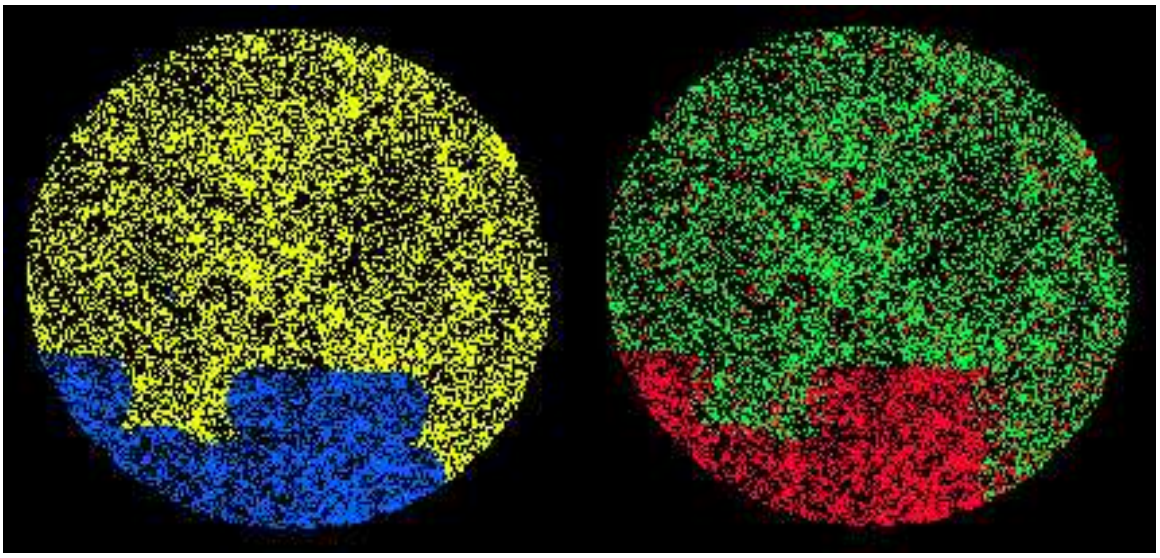

Figure 6.2: Two views of the circular projection of the cell from Figure 6.1. On the left, 15,381 MGS targets within the spectroscopic footprint are colored in yellow while the 5125 outside are colored in blue. On the right, MGS galaxies are colored in green while MGS objects are colored in red. We refer to objects within the contiguous green region as *interspersed objects* and those outside it as *dark objects*.

similar study for 11 photo-z codes.

Rather than attempt a thorough comparison of all methods in the literature, we focus on a subset of nine fundamental techniques. We do this with the understanding that the best counting methods may possibly be excluded from this analysis. This is by no means a judgment on these methods' merits, but merely a necessity to constrain the scope of our investigation. The work that follows is designed as an initial study from which future analyses can be extended.

6.1.1 Galaxies with Redshifts

Counting the number of MGS galaxies in each cell is straightforward. Each galaxy possesses a well-calibrated angular position and high-quality redshift. These are mapped to comoving radial distances $\chi(z)$ through equation (A.7) to establish their three-dimensional positions. If (X_i, Y_i, Z_i) and (X_c, Y_c, Z_c) are the comoving coordinates of the i^{th} galaxy and center of a sphere respectively, the i^{th} galaxy resides within the cell if

$$(X_i - X_c)^2 + (Y_i - Y_c)^2 + (Z_i - Z_c)^2 \leq r_c^2, \tag{6.1}$$

where r_c is the radius of the cell.

6.1.2 Discrete Counting

Discrete counting methods assign each object a single redshift. This approach offers the best upside. Under optimal (though admittedly unlikely) conditions, discrete counting can be exactly right, something scaling and probabilistic smearing methods cannot offer.

The errors induced by discrete counting of a single object are limited to at most two non-overlapping cells — the one it is estimated to reside inside, and the one it is actually inside. In this way, the negative impact of discrete counting is limited in the number of cells it can affect, though the absolute errors in those cells are potentially much larger than with other methods.

Ignore: The simplest of the nine methods, "ignore" simply disregards every object during the counting process. This method will systematically undercount the number of targets in each cell. For cells with large angular projections, ignoring all objects can discount a significant number of targets, leading to large errors. This method should improve as cell size shrinks or the number of cells per unit redshift increases. Either scenario will decrease the number of objects intersecting each cell's line-of-sight.

Template and Training Set Photometric Redshifts: Each object is assigned a redshift equal to its photometric redshift. This implicitly correlates radial distance with an object's brightness and color profile, as described in §2.1.4. Two types of photometric redshifts will be tested. The first are template based photo-z's, which utilize an object's spectral energy distribution. These will hereafter be abbreviated "SED photo-z's". The second are training set, or ANN, photo-z's. The SDSS database offers two varieties — D1 and CC2. We utilize D1 photo-z's since these have been shown to display better performance at brighter magnitudes. They are drawn from CAS table `Photoz2`.

Nearest Neighbor Method: Each object is assigned the redshift of the galaxy that lies the smallest angular separation away (e.g. Zehavi et al., 2002, 2005; Berlind et al., 2006). Nearest neighbor was an early solution for handling fiber collided galaxies in SDSS-I/II (Zehavi et al., 2002, 2005, 2011) down to $\sim 0.1\,h^{-1}\mathrm{Mpc}$. It has also been used to identify galaxy groups and clusters (Berlind et al., 2006). This method works better when angular

separation is small, though it has had difficulty below the fiber collision scale and when constructing the redshift-space correlation function. Nearest neighbor is weaker at large redshifts, where a given angular separation implies a larger physical separation. It is expected to be less effective in dark regions and external regions.

6.1.3 Scaling

Rather than directly approximate their redshifts, the *scaling method* accounts for the presence of objects by upweighting the count of galaxies in cells. The scaling mechanism works differently for *interspersed objects* and *dark objects*.

Consider first the case in which dark objects lie in well-defined dark regions formed by the intersection of a cell's circular projection with the spectroscopic and photometric footprints. Assume dark regions contain only objects, and that the remainder of the cell contains only galaxies. (Recall that the volume of the cell within those footprints are β_{spec} and β_{PS}, respectively. An example was provided in Figure 6.2). If n_g is the number of galaxies in the cell, the scaling method approximates the number of dark objects within its volume to be

$$n_d = \left(\frac{V_d}{V_g}\right) n_g = \left(\frac{\beta_{PS} - \beta_{spec}}{\beta_{spec}}\right) n_g, \tag{6.2}$$

where V_d is the volume of the cell intersected by the dark region and V_g is the volume of the cell inside the photometric footprint. The total number of targets n_t approximated to lie

within the cell becomes

$$n_t = \left(\frac{\beta_{PS}}{\beta_{spec}} \right) n_g, \tag{6.3}$$

where $c = \beta_{PS}/\beta_{spec}$ is the "scaling factor". The scaling method, as represented through equation (6.3), can only be employed when the cells contain distinct, contiguous, spectroscopically unsampled regions, i.e. when β_{PS} and β_{spec} are clearly specified. Situations like this are common when cells are placed near the edges of a spectroscopic footprint, or when masks are introduced.

Assuming one's survey is photometrically complete, the scaling method should only be employed if a cell's dark region actually contains dark objects. If no dark objects are present, the approximate number of targets in the cell's dark volume is trivially set to zero. In this way, each dark object is not counted individually, but rather, acts as a binary switch for whether the scaling method will be executed or not.

A disadvantage of this method is that useful information — the number of dark objects in a cell's dark region — is essentially discarded. A single dark object that intersects numerous cells along its line-of-sight can potentially contribute to an aggregate number count totaling more (or less) than one distinct object. This almost guarantees that the total number count of dark objects will not be conserved, an issue that resurfaces with the probabilistic smearing methods introduced in §6.1.4.

Because scaling relies heavily on the approximated number densities of galaxies in cells, it should be most effective when the dark regions' total area is small relative to the

cell's projection. It is better suited for cells with large projections since the number densities inside and outside the dark regions are more likely to be similar and less likely to vary due to fluctuations in large scale structure.

Next, consider the case in which *interspersed objects* are distributed among interspersed galaxies within a cell's interspersed region. Because there are no clearly delineated areas containing only objects or only galaxies, equation (6.3) is inapplicable. Instead, the scaling factor c can be calculated using the spectroscopic completeness of the interspersed region. To first approximation $c = 1/f$ where f is the spectroscopic completeness of the survey.

However, as seen in §5.3.1 spectroscopic completeness is anisotropic and, in some cases, extremely so. While the footprint corrections helped equalize the completeness across the spectroscopic footprint as a whole, there are still localized clusters where a disproportionately small (or large) number of MGS targets were assigned fibers.

We therefore propose a direction-dependent spectroscopic completeness factor

$$c = \begin{cases} 1 & \text{if } N_g = 0 \text{ and } n_{io} = 0 \\ N_g/(N_g + n_{io}) & \text{otherwise} \end{cases}, \tag{6.4}$$

where N_g and n_{io} are the numbers of *galaxies* and *interspersed objects* whose projections intersect the cell's interspersed region. The total approximated number count of targets within the interspersed region is then

$$n_t = \begin{cases} 0 & \text{if } c = 0 \\ \\ n_{ig}/c & \text{otherwise} \end{cases}, \tag{6.5}$$

where n_{ig} is the number of *interspersed galaxies* inside the cell's volume as given by equation (6.1). The distributions of completeness factors c are presented in Figure 6.3.

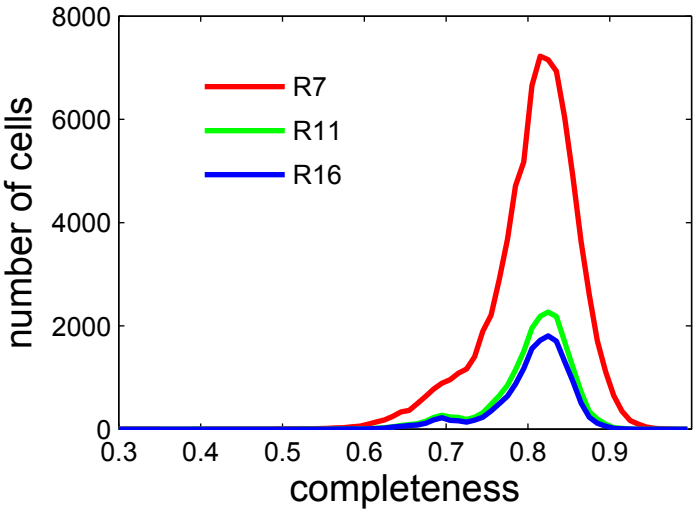

Figure 6.3: Distribution of direction-dependent completeness factors c. Factors are counted in bins of width $\Delta c = 0.01$.

Unlike with dark regions, the number of interspersed objects n_{io} acts as more than a binary switch. By helping to parameterize c, this scaling approach introduces a proximity bias. If a cell contains no galaxies within its volume, it cannot contain any objects either. This effectively limits object counts to volumes where galaxies already exist.

Interspersed scaling works best when the radial distribution of objects is the same as that of galaxies. While this is certainly not the case in all directions, it should hold reasonably well along restricted lines-of-sight. The circular projections of R7, R11, and R16 cells,

especially those at large redshifts, fall squarely into this category.

A final scaling option would ignore MGS objects entirely and renormalize the selection function to only account for the presence of MGS galaxies. This tactic would adjust the expected number of galaxies downward, similar to how the scaling method would adjust total number count upward. In either case their ratio, which determines the value of the overdensity δ, would remain the same, save some small differences induced by anisotropies in the spectroscopic completeness. Because the differences between scaling number count versus expected number is so slight, the latter option is not examined any further in this analysis.

6.1.4 Probabilistic Smearing

Any method for estimating a galaxy's depth is an attempt to correlate its radial distance with other properties like position, brightness, color, or spectral characteristics. None of these properties supply a deterministic link to depth, but rather, offer a probability distribution that the galaxy lies within a given redshift range. Even a galaxy's spectral classification, one of the measurements most tightly correlated to depth, is subject to uncertainties due to effects like peculiar velocities and errors in the cosmological model.

The probabilistic smearing methods introduced in this section operate not by assigning an object any particular redshift, but by reporting the probabilities that the object lies in each cell along the line-of-sight. Under this methodology a single object count is *smeared* among multiple cells, with each cell i along the line-of-sight receiving a partial count n_i

such that $\sum_i n_i \leq 1$.

Unless the probability distribution $p(z)$ is very sharply peaked within a cell, the number count will split itself among multiple cells and the "empty" areas in between. It may also extend beyond the redshift boundaries of the survey itself. Consequently, in most realistic scenarios $\sum_i n_i$ will be less than 1 and some of the count will be "lost."

The probability that an object lies within the boundaries of a cell depends upon two factors — the entry and exit points of the line-of-sight chord through the cell and the shape of the density function itself. A chord that passes directly through the center of a cell will contribute a higher number count than one that glances its edge even for cells at the same redshift.

Letting z_l^{ij} and z_u^{ij} represent the entry and exit redshifts for object j's line-of-sight chord through cell i (see equation (E.9)), the partial count contributed by the object to the cell is

$$n_{ij} = \begin{cases} w_j \int_{z_l^{ij}}^{z_u^{ij}} p_j(z)\, dz & \text{if } \hat{n}_i \cdot \hat{x}_j > c_i \\ 0 & \text{otherwise} \end{cases} . \tag{6.6}$$

The total number count in cell i will be $\sum_j n_{ij}$. The function $p_j(z)$ is normalized such that its integral over the redshift range within which the object is constrained exist equals unity. The weighting factor w_j reflects how heavily to count one object relative to the others.

For the purposes described in this dissertation, $w_j = 1\ \forall j$ as all objects are weighted equally. It is possible, however, to develop hybrid criteria through which a fraction $w < 1$ of an object is smeared through one method, while the remainder $1 - w$ is counted via

another. This more sophisticated approach embodies a dual probability system — one in which the full set of an object's properties is used to nominate multiple counting methods, each applied with a different weight. We discuss this approach for completeness, but do not further explore its usage in these pages.

Because probabilistic smearing deposits partial counts in cells, it almost certainly will not yield a result that is exactly correct. This is a critical distinction from discrete counting methods. The hope is that if the probability models are accurate, the combination of smeared counts from multiple objects will better approximate the true distribution of targets than other counting alternatives. As will be shown, later there is good reason to suspect this is the case, at least for the survey as a whole.

Below, four probabilistic smearing methods are introduced and explained. Each distributes galaxy counts in cells using equation (6.6). They differ only in the probability density functions $p(z)$ used to do so.

Selection Function Smearing: The method of *selection function smearing* takes $p_j(z) = p_{exp}(z)\ \forall j$, where p_{exp} is the distribution of MGS targets expected in the absence of clustering (see equation (2.14)). In a sense, this can be considered a default case for smearing in that no object's property is utilized other than the fact that it is drawn from the same distribution as the MGS galaxies. While smearing via the selection function guarantees that no clustering statistics beyond homogeneity will be reported, it may serve as a useful tool to "fill in the gap" left by an object's absent redshift.

Selection function smearing follows the guiding principle of Guo et al. (2012)'s method of recovering the 2PCF in light of collided objects. Guo's analysis relies upon the assumption that the collided objects have the same distribution as the known galaxies. In their method, the radial location of an object is not correlated to any of its own measurable characteristics (e.g. color, photo-z, proximity to nearest neighbor), but to the simple fact that it is a member of a survey assumed to have the same radial distribution.

Two-Point Correlation Function Smearing: The *2PCF smearing* method quantifies the probability that an object lies between two redshifts by using the two-point correlation function $\xi(r)$. By necessity, the 2PCF considers galaxies as a pair. We chose to partner each object of unknown redshift with its nearest galaxy neighbor of known redshift, since this is the galaxy with which the object will have the largest spatial correlation.

From equation (3.5), the joint probability $dP(\boldsymbol{r})$ of finding galaxies in each of two separate volume elements dV_1 and dV_2 separated by a vector $\boldsymbol{r}$ is $dP(\boldsymbol{r}) = n^2(1 + \xi(\boldsymbol{r}))\ dV_1\, dV_2$, where n is the background density that would exist if the Universe were perfectly homogeneous. If one galaxy's position is fixed, the probability of finding another galaxy a distance r away is found by marginalizing over one of the volume elements. The Universe is assumed isotropic and the directional dependence on $\boldsymbol{r}$ is dropped,

$$dP(r) \propto n(1 + \xi(r))\ dV. \tag{6.7}$$

The differential volume element can be written $dV = A\, dz$ where A is an arbitrarily-

sized cross sectional area. The background density of galaxies is independent of direction but in a magnitude-limited survey it is redshift-dependent. Therefore,

$$dP(r) = Cp_{exp}(z)(1 + \xi(r))\ dz, \tag{6.8}$$

where C is a normalization constant. If the object of unknown depth is constrained to exist between fixed redshift limits z_i and z_f then the normalization is fixed such that

$$C = \left[\int_{z_i}^{z_f} p_{exp}(z)\, dz + \int_{z_i}^{z_f} p_{exp}(z)\, \xi(r)\, dz\right]^{-1}. \tag{6.9}$$

When two galaxies' positions are uncorrelated the second integral in equation (6.9) vanishes and this method reduces to selection function smearing. When r is small, $\xi(r)$ dominates and a galaxy's depth is mostly constrained by the two-point correlation function. In this way, 2PCF smearing is a combination of selection function smearing and a probabilistic version of the nearest neighbor method.

Integration is done numerically. Redshifts between z_l^{ij} and z_u^{ij} are computed on a z-grid with resolution $dz = 10^{-5}$. At each grid location, a conversion to comoving depth is performed and $p_{exp}(z)$ is evaluated. The distances r between the nearest neighbor and each comoving grid point are found and the values of $\xi(r)$ are subsequently interpolated. The integration then proceeds as normal. The final form of the probability density function is quite weak and introduces only a minor perturbation to the selection function.

$$p(z) = Cp_{exp}(z)(1 + \xi(r)) \,. \tag{6.10}$$

Figure 6.4 displays three sample 2PCF smearing density functions. Each possesses the shape of the selection function with a spike at the redshift of its nearest galaxy neighbor. As the nearest neighbor separation decreases, the height of the spike increases, indicating an increased likelihood that the object resides close to its nearest neighbor. Conversely, the probabilities that the object resides at all other redshifts decrease accordingly. In the case of the largest angular separation of 2096 arcseconds, the nearest neighbor correlation is quite weak and introduces only a minor perturbation to the selection function.

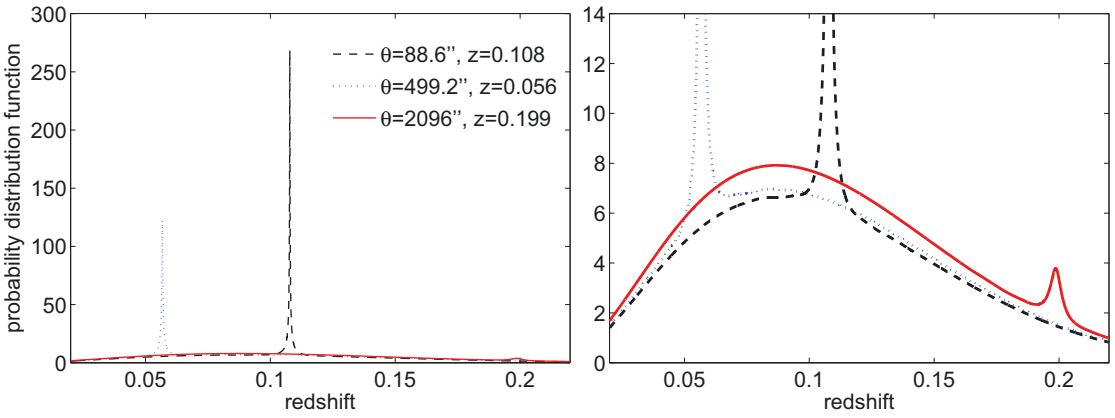

Figure 6.4: The probability distribution functions of three object/nearest-galaxy-neighbor pairs are displayed. The functions are selected to span a wide range of nearest neighbor redshifts and angular separations (given in arcseconds). All three functions are normalized to 1 between $z = 0.02$ and $z = 0.22$.

Photometric Redshift Smearing: The final two smearing methods correlate depth to

color and luminosity through the SED and D1 photometric redshifts and their associated errors. As first explained in §2.1.4, photometric redshifts are Bayesian reconstructions that may be interpreted as Gaussian probability functions. In these cases, $p(z) = C\, g(z)$, where $g(z)$ follows a Gaussian distribution with a mean μ_z equal to the photometric redshift, and a variance σ_z^2 equal to the square of the error reported in the SDSS database,

$$g(z) = \exp\left[-\frac{(z-\mu_z)^2}{2\sigma_z^2}\right]. \tag{6.11}$$

The normalization constant is again set by specifying the redshift limits within which the object is constrained to exist,

$$C = \left(\int_{z_i}^{z_f} g(z)\, dz\right)^{-1}. \tag{6.12}$$

Redshift limits $z_i = 0.02$ and $z_f = 0.30$ are selected to match the range of MGS galaxies.

The assumption of photo-z Gaussianity has precedent. Balogh et al. (2014) model photometric redshifts with a Gaussian distribution. They and others (Ilbert et al., 2009; George et al., 2011) have shown that integrating it is an effective way to count galaxies within a redshift range. López-Sanjuan et al. (2010) have used the Gaussian model to identify close galaxy pairs.

Several studies have concluded that the best way to utilize photo-z's is probabilistically (e.g. Fernández-Soto et al., 2002; Cunha et al., 2009; Wittman, 2009; Myers et al., 2009; Carrasco Kind & Brunner, 2014). A plethora of algorithms exists to do so. Carrasco Kind

& Brunner (2014) combine different models into a stronger estimator through a Bayesian framework. Some groups get probability distributions directly from the photometric templates. More sophisticated analysis represents the total PDF as a combination of red and blue templates.

The photo-z smearing methods implicitly ignore any spatial correlations that might be present. This sacrifice becomes less of a problem at high redshifts where galaxies are sparse and their physical separations start to exceed the spatial correlation radius. At this point, it becomes statistically unlikely that a galaxy is spatially correlated with its nearest neighbor, and the photometric redshift — which will have the narrower core — begins to carry more information.

A preliminary comparison of these counting methods is presented in Figure 6.5. The true distribution of pristine MGS galaxies (as measured through spectroscopic redshift), is plotted along with those of the selection function and photo-z smearing methods. Note that there is no difference between the suitably normalized selection function and the aggregate result of selection function smearing.

It is immediately obvious that none of the methods are capable of reproducing the true galaxy distribution on small scales. The errors inherent to each method are larger than this level of detail, which reinforces the idea than none are suitable on their own for conducting high-precision redshift surveys. The net result is a smoothing effect everywhere except the peak of the discrete photo-z distribution.

Of the three methods, photo-z smearing is best able to match the true redshift distribu-

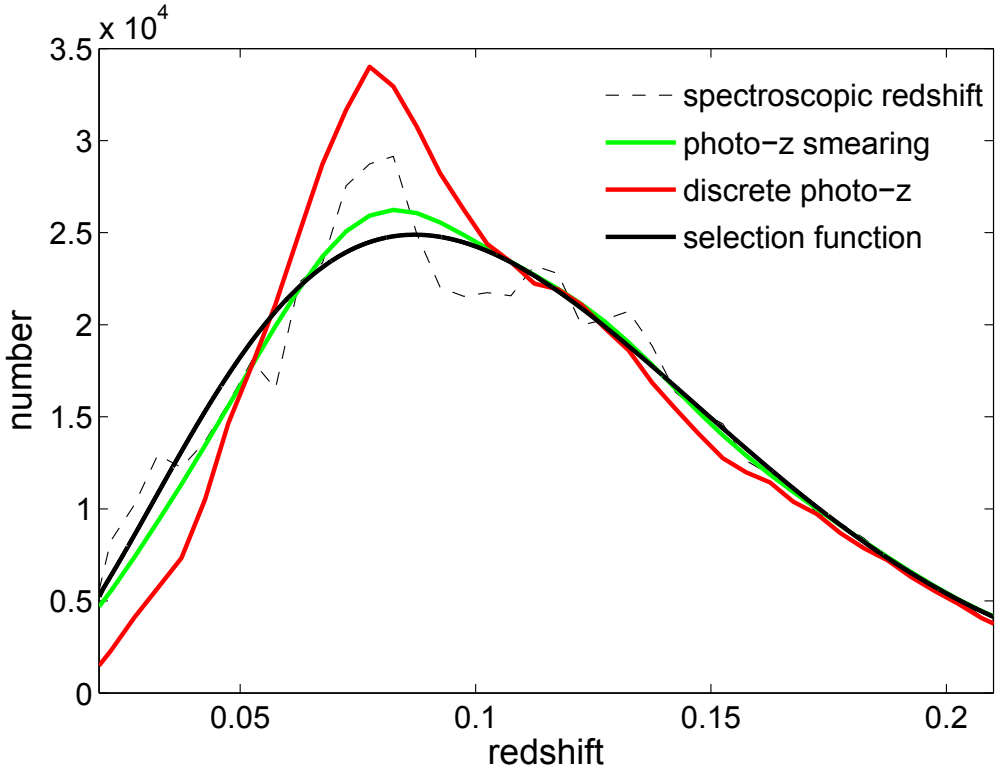

Figure 6.5: The redshift distribution of DR8 pristine MGS galaxies is presented alongside the distributions approximated from selection function smearing, SED photo-z smearing, and discrete SED photo-z counting. The D1 photo-z plots overlap the SED plots almost exactly and are not included. Galaxies are bounded by the range $z \in [0.006, 0.956]$ and are counted in bins of size $\Delta z = 0.005$. The boundaries of the photo-z smearing Gaussian integrals are $z_i = 0$ and $z_f = 1$.

tion of MGS galaxies for $z > 0.04$. The fact that this methods induces no large, systematic shift from the true distribution suggests that the photometric redshifts (and their errors) contain more information about large scale structure than does the selection function alone.[2] Discrete photometric redshifts offer the worst performance, indicating that a photo-z's uncertainty carries useful information that should not be discounted.

[2]In addition to galaxy counting, photo-z smearing has a number of other applications, like understanding the statistical distribution of a subsample of galaxies, perhaps due to a color or magnitude cut. This approach might be used to determine whether galaxy targets are behind a lensing object, for example.

6.2 An Empirical Signal Correlation Function

The strength of the 2PCF smearing method depends upon how well the correlation function represents the data to which it is applied. Up until this point, we have utilized a redshift-space 2PCF calculated using a fiducial power spectrum and parameterized with a fixed cosmological model (see §3.4). In this section, we calculate an empirical two-point correlation function $\xi(r)$ using only the MGS galaxies pulled from the northern hemisphere of the DR6 improved spectroscopic footprint.

There are a few reasons for making this change. First, we want to ensure that the correlation function used to test the 2PCF smearing method accurately reflects the distribution of galaxies being counted. In removing all MGS objects from the simulations, we have changed the distribution of galaxies $P_{fid}(k)$ is supposed to represent. Aligning the real and modeled correlation functions reduces a discrepancy that could negatively bias our results. One downside of this approach, however, is that redshift-space distortions will no longer be incorporated into $\mathbf{\Sigma}_{\kappa}$.

The empirical 2PCF is calculated using the pair counting method of Landy & Szalay (1993). The distances between all galaxy pairs DD are measured, binned in bins of width $r \pm dr$, and counted. The same is done for arbitrarily dense random points RR that are distributed uniformly as a function of angle within the improved spectroscopic footprint and radially according to $p_{exp}(z)$ from equation (2.14).

According to Landy and Szalay, the estimator

$$\xi(r) = \begin{cases} \left(\frac{DD}{N_{DD}} - \frac{2DR}{N_{DR}} + \frac{RR}{N_{RR}} \right) / \frac{RR}{N_{RR}} & r < 89.575\,h^{-1}\mathrm{Mpc} \\ 0 & \text{otherwise} \end{cases}, \tag{6.13}$$

minimizes the variance in $\xi(r)$ to the Poisson level, where DR are the pair counts of the cross-correlated galaxies and randoms, and which are introduced to account for survey edge effects. (The dependence of DD, RR, and DR on r in equation (6.13) is implied for notational simplicity.)

The accuracy of the Landy & Szalay estimator depends upon the number of random points N_R used. We take $N_R = 10N$ where N is the number of MGS galaxies in the northern hemisphere of the spectroscopic footprint. This gives way to $N_{RR} = N_R(N_R - 1)/2$ unique pairs of random points, $N_{DD} = N(N-1)/2$ pairs of galaxies and $N_{DR} = NN_R$ galaxy/random pairs. The cosmological principle is in effect beyond $r \geq 100\,h^{-1}\mathrm{Mpc}$, meaning homogeneity and isotropy are typically maintained beyond these scales. In practice we find that $\xi(r)$ first runs negative at $r = 89.575\,h^{-1}\mathrm{Mpc}$, so the correlation function is set to zero beyond that point.

After convolving equation (6.13) with the spherical windows from equation (4.3), the correlation functions in Figure 6.6 result. The elements of the empirical signal covariance matrix $\mathbf{\Sigma}_\kappa$ are assembled in the manner described in §4.4 with the one exception that the distances between cells are calculated using Euclidean, rather than the Liske, geometry to better match the way in which $\xi(r)$ is evaluated.

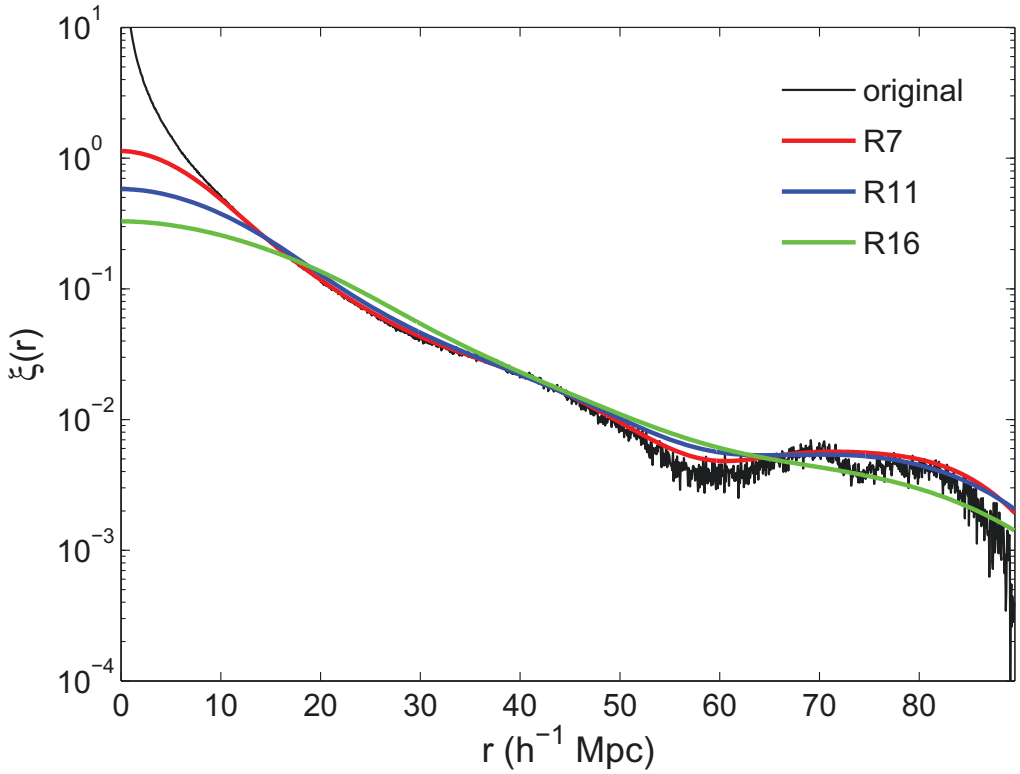

Figure 6.6: Empirical two-point correlation function as calculated using MGS galaxies from the northern hemisphere of the improved spectroscopic footprint. The ratio from equation (6.13) is displayed (*black*) along with its convolutions with spherical window functions of radii 7, 11, and 16 $h^{-1}\mathrm{Mpc}$ (*red*, *blue*, *green*). Original $\xi(r)$ is binned in bins of width $2\,dr = 0.05\,h^{-1}\mathrm{Mpc}$.

6.3 Testing Under Three Scenarios

In this section we simulate the presence of MGS *galaxies* and *objects* in three different region types — interspersed regions, dark regions, and external regions — in order to test which of the nine counting techniques described in §6.1 are most effective under various conditions.

In lieu of constructing full MGS mock catalogs, we generate *galaxy* and *object* realizations by randomly dividing the 474,651 MGS galaxies within the improved spectroscopic

footprint into two sets — those that retain their redshift information, and those that do not. Because the true redshifts of MGS galaxies are known, their number in each cell is fixed. This "ground truth" may be compared against counts estimated through each of the nine counting techniques to draw conclusions about their efficacy.

Using the real set of MGS galaxies to create *galaxy/object* simulations offers some advantages. Randomizing which MGS galaxies retain their redshifts allows for the rapid generation of multiple Universes in which the constituent targets share the same properties (both spatial and photometric) as those we are trying to count. In this way, difficult to model correlations between spatial clustering and photometric redshift do not need to be discovered before simulations can commence. This increases the likelihood that conclusions drawn from simulations will be applicable to the full set of MGS targets.

A downside of this approach is that the spatial distribution of the MGS galaxies is static. This means that conclusions drawn from their realizations will inevitably be weighted towards the SDSS's local sampling of Universe, and may not be as robust on average. Also, the full MGS within the spectroscopic footprint contains about 25% more targets than does the set of MGS galaxies from which the simulations are drawn. Consequently, the conclusions reached in this chapter will be for a somewhat sparser Universe than reality. We do not attempt to account for this difference beyond advising that even if conclusions are imperfect, they should still be able to offer informative first-order principles to be carried forth into even more robust analyses.

To assess the effectiveness of each counting technique as a function of redshift, we in-

troduce an *error metric* $\epsilon^{(\,)}$ that quantifies each method's impact on galaxy number count n, overdensity δ, and overdensity squared δ^2 by measuring the average size of the discrepancy between the truth and each realizations τ. Let n_i equal the true galaxy count in cell i. Let $n_i^{(\tau)}$ equal the count in cell i during realization τ using one's method of choice. We report the average discrepancy over K realizations,

$$\Delta n(z_i) \equiv \sum_{\tau=1}^{K} |n_i - n_i^{(\tau)}|/K, \tag{6.14}$$

where z_i is the redshift of cell i. We split redshift-space into a discrete number of bins and let the boundaries of the j^{th} bin be z_j and z_{j+1}. The deviation is ultimately reported as an average over all the cells in each redshift bin,

$$\epsilon_j^{(n)} = \langle \Delta n(z_i) \rangle, \; z_j \leq z_i < z_{j+1}. \tag{6.15}$$

The error bars on $\epsilon_j^{(n)}$ are reported as the 1σ standard deviation of $\Delta n(z_i)$ for all cells within the corresponding redshift range. A similar statistic is applied to the overdensities and overdensities squared by replacing $|n_i - n_i^{(\tau)}|$ in equation (6.14) by $|\delta_i - \delta_i^{(\tau)}|$ and $|\delta_i^2 - \delta_i^{2(\tau)}|$ respectively.

6.3.1 Interspersed Regions

One realization of an interspersed region is generated by randomly stripping 20% of MGS galaxies of their redshifts and turning them into *objects*. Galaxy counts are approxi-

mated and the process repeats until 50,000 counts are gathered within each redshift range. The density of MGS galaxies is large enough that we did not encounter any situations in which a cell's circular projection did not contain at least one *object*.

An *object*'s nearest neighbor is defined to be the *galaxy* at the smallest angular separation. To speed up testing, the ten nearest neighbors for each MGS galaxy are precomputed. In the rare event that an *object*'s ten nearest angular neighbors are also stripped of their redshift information during a realization, the tenth of those assumes the role of nearest neighbor.

The counting results for the R7, R11, and R16 cases are presented in Figures 6.7, 6.8, and 6.9 respectively. All measurement types for all cell sizes share a couple common features. Among the methods tested, ignoring *objects* produces the worst results at low redshifts, where galaxies are plentiful. This fact is increasingly apparent as cell size increases. However, ignoring objects often stands as the best alternative at high redshifts where galaxies are scarce. That is, so few galaxies actually reside at high redshifts that assuming no objects lie there is actually a sound strategy.

Among the photo-z's and probabilistic smearing methods, no clear favorite emerges. For most measurement types and redshifts, these methods tend to lie within 2σ of each other. A notable exception is the R16 case, where for $z > 0.22$ the photometric smearing methods fail spectacularly. This version of probabalistic smearing tends to overestimate the probability of finding objects at high redshifts, a conclusion that re-emerges in other tests that follow.

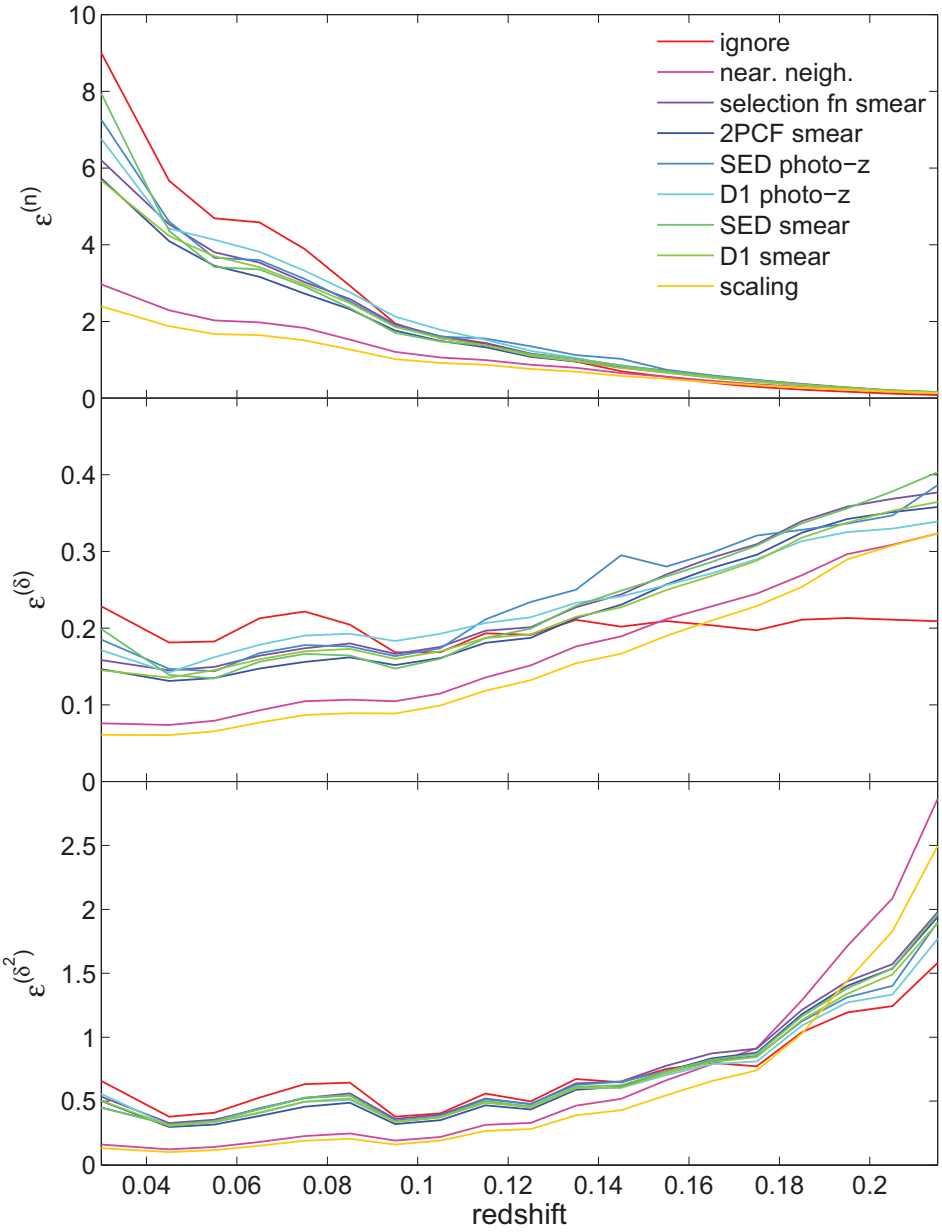

Figure 6.7: Error metrics from equation (6.15) for number counts n, overdensities δ, and overdensities squared δ^2 for R7 cells in interspersed regions. Error metric values are averaged over redshift bins of width $z = 0.01$. Error bars are omitted for visual clarity here, but are available in text files online. Uncertainties for select counting methods and comparisons are also plotted in Figures 6.10, 6.33, and 6.34.

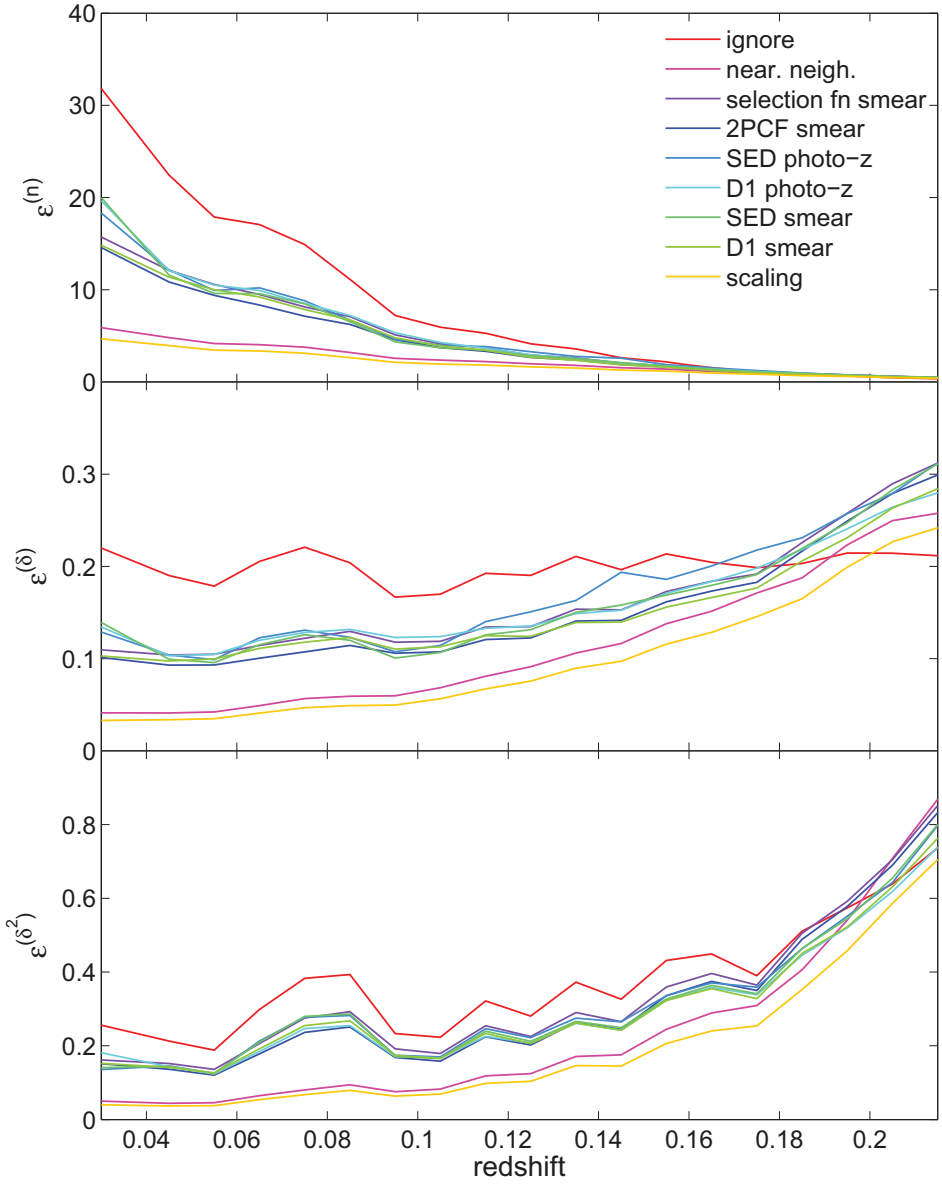

Figure 6.8: Error metric from equation (6.15) for number counts n, overdensities δ and overdensities squared δ^2 for R11 cells in interspersed regions. Error metric values are averaged over redshift bins of width $z = 0.01$. Error bars are omitted for visual clarity here, but are available text files online. Uncertainties for select counting methods and comparisons are also plotted in Figures 6.10, 6.33, and 6.34.

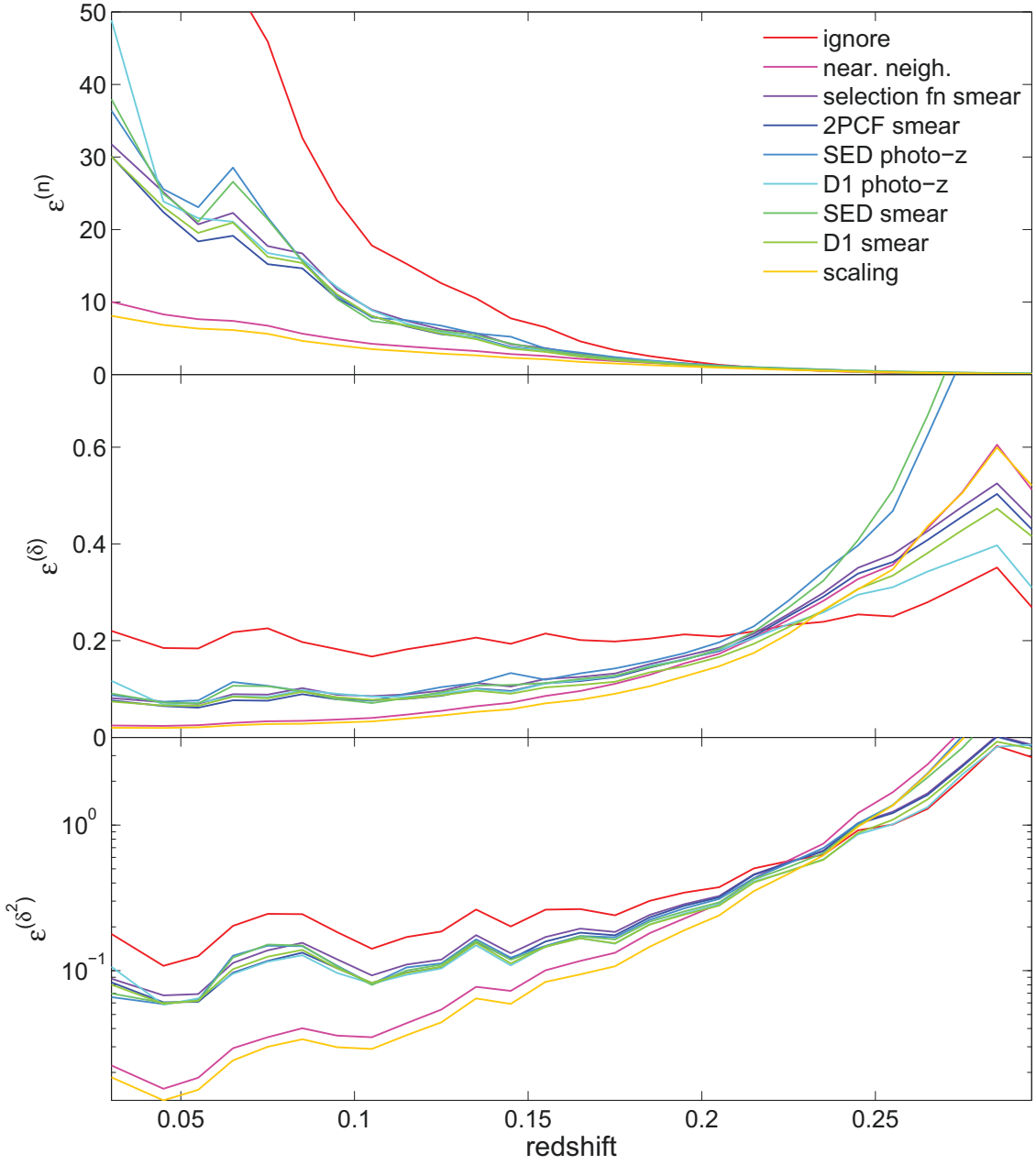

Figure 6.9: Error metric from equation (6.15) for number counts n, overdensities δ and overdensities squared δ^2 for R16 cells in interspersed regions. Error metric values are averaged over redshift bins of width $z = 0.01$. Error bars are omitted for visual clarity here, but are available in text files online. Uncertainties for select counting methods and comparisons are also plotted in Figures 6.10, 6.33, and 6.34.

The optimal counting techniques for interspersed regions are presented in Table 6.1. Scaling is the preferred method at low and intermediate redshifts. As a practical matter, this implies that the optimal strategy (among the methods tested) for handling fiber collided galaxies is to use *only* MGS galaxies to both: 1) normalize the selection function *and* 2) count targets in cells. At high redshifts, ignoring objects is preferred. The redshift at which this transition occurs increases with cell size.

The popular nearest neighbor method is nonoptimal in all situations. As Figure 6.10 illustrates, that result is statistically significant. This is especially true for $\Delta\epsilon^{(n)}$ and $\Delta\epsilon^{(\delta)}$ when $z \lesssim 0.16$. The preference for using scaling over the nearest neighbor method when estimating n is strongest when either cell radius or $\langle n \rangle$ is large. That trend is reversed for δ and δ^2, where the ratio of number count to $\langle n \rangle$ plays the larger role. The superiority of scaling over nearest neighbor only loses 2σ statistical significance at the highest redshifts. However, this lack of significance is largely moot since ignoring objects dominates both methods at high z.

6.3.2 Dark Regions

When cells are placed within the SDSS footprint using the HCP arrangement, some of their volumes will inevitably reach into areas outside the spectroscopic footprint but within the photometric footprint. We refer to these areas as *dark regions*. We start this section by explaining how to simulate dark regions with the same statistical properties as those actually present within the DR6 MGS survey. We then use these regions to test the

	R7			R11			R16		
z	n	δ	δ^2	n	δ	δ^2	n	δ	δ^2
0.030	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.045	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.055	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.065	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.075	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.085	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.095	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.105	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.115	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.125	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.135	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.145	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.155	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.165	ignore	ignore	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.175	ignore	ignore	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.185	ignore	ignore	scaling	scaling	scaling	scaling	scaling	scaling	scaling
0.195	ignore	ignore	ignore	scaling	scaling	scaling	scaling	scaling	scaling
0.205	ignore	ignore	ignore	ignore	ignore	scaling	scaling	scaling	scaling
0.215	ignore	ignore	ignore	ignore	ignore	scaling	scaling	scaling	scaling
0.225							scaling	scaling	scaling
0.235							ignore	ignore	D1sm.
0.245							ignore	ignore	D1
0.255							ignore	ignore	ignore
0.265							ignore	ignore	ignore
0.275							ignore	ignore	ignore
0.285							ignore	ignore	D1
0.295							ignore	ignore	ignore

Table 6.1: Optimal counting techniques for interspersed MGS objects as a function of redshift.

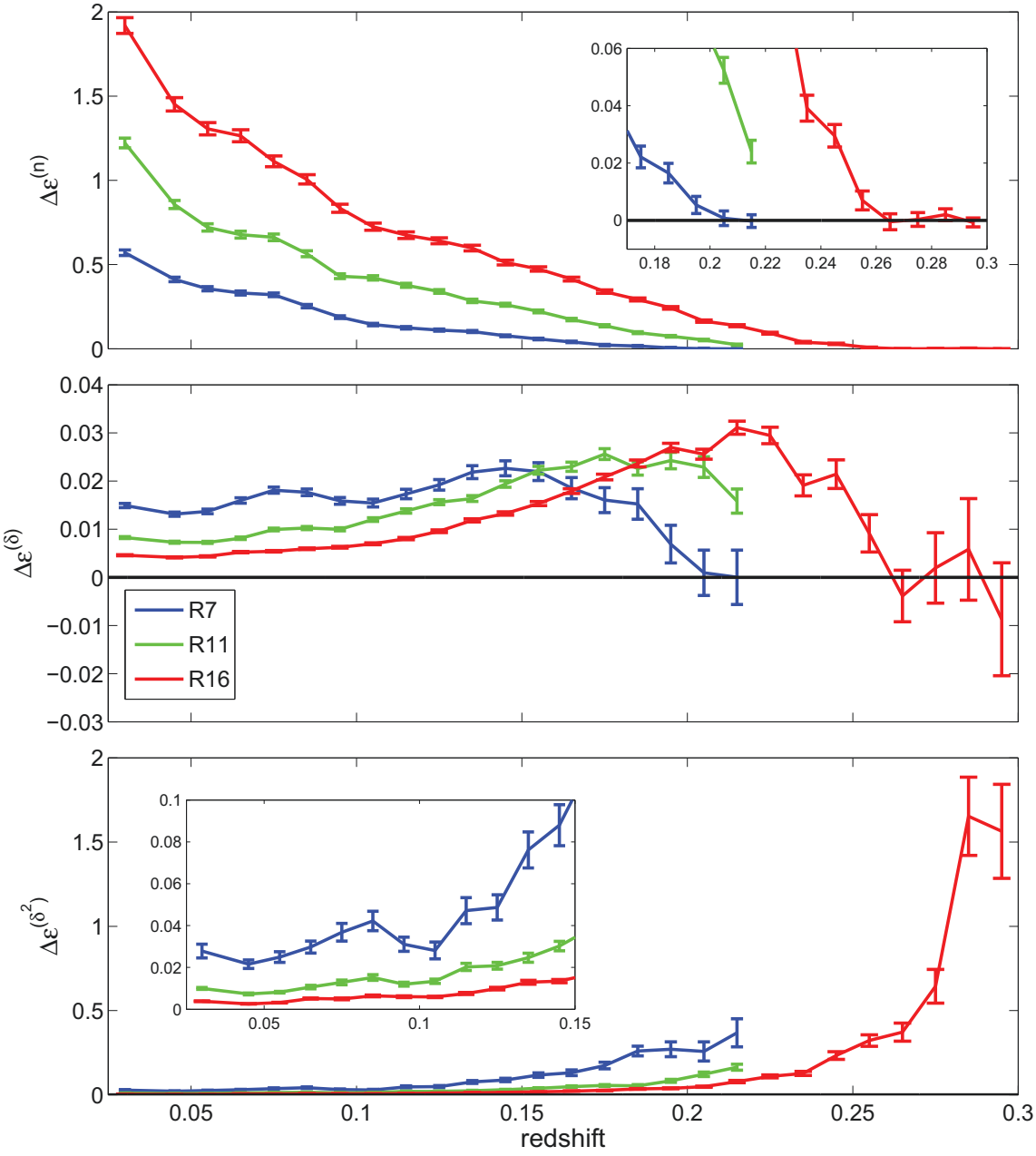

Figure 6.10: A comparison between error metrics for the scaling and nearest neighbor methods. The vertical axis for each subplot represents $\epsilon^{(\,)}$ for nearest neighbor minus $\epsilon^{(\,)}$ for scaling. Redshifts for which $\Delta\epsilon^{(\,)} > 0$ indicate that scaling is preferable to nearest neighbor at those locations.

counting methods in a similar manner to the interspersed regions.

6.3.2.1 Simulating Dark Regions

To simulate the effects of dark regions, we must construct shapes that mimic their properties. Simple candidates include circles, rectangular patches, and flattened arcs. Various factors must be considered. Should their areas match those of the true dark regions, or is it more important for the distribution of objects' nearest neighbor distances to be the same? Should they contain the same number of objects as the true dark region, or is it more important to have perimeters of equal length? Is some combination of these qualifications called for?

Whichever choice is made, it is clear that the simulated dark regions must be a function of redshift. To see why, consider a dark region like that in Figure 6.1 that covers a relatively large area. This low redshift R16 cell contains a dark region that reaches further beyond the edge of the spectroscopic footprint than smaller, higher redshift cells can for a fixed β_{spec}. In fact, the area of this dark region can exceed the total area of higher redshift cells, meaning all interior targets within those cells could be stripped of their redshifts during simulations. Because the true dark regions in our survey never exceed 38% of a cell's volume, simulated dark regions must obey a similar constraint.

We avoid this effect by creating distinct sets of "dark region random variables" — those that populate a distribution from which representative dark regions may be drawn — for each of a handful of "redshift slices". We determine the properties of dark regions within

each redshift range $z_i < z < z_{i+1}$ to parameterize the distributions, then apply simulated dark regions only to cells within the same range.

Edge effects pose a concern. When testing the nearest neighbor and 2PCF smearing methods, it is important that the distribution of the dark objects' nearest neighbor distances matches reality. The majority of cells for which $\beta_{spec} < 1$ lie near the edges of the survey. Their dark regions of area A are typically adjacent to galaxies (i.e. nearest neighbor candidates) only on the side directed towards the spectroscopic footprint. If our dark regions are simulated with, say, a circular patch of area A, the average nearest neighbor distances between dark objects inside the patch and nearest neighbors outside the patch will likely be smaller than needed to adequately mimic the properties of the true dark regions.

The better solution is to create dark region random variables that *perfectly match* the shapes of the true dark regions. Recall that true dark regions are created by intersecting a cell's circular projection (given by halfspace constraints $[\boldsymbol{r}_c, c]$) with the improved spectroscopic footprint. The cell's center may be repositioned to point $\boldsymbol{r}_p$ using a rotation matrix $\mathbf{M}$ such that $\boldsymbol{r}_p = \mathbf{M}\boldsymbol{r}_c$. Multiplying each of the spectroscopic footprint's halfspace constraints by $\mathbf{M}$ realigns the dark region with its cell. If $\boldsymbol{r}_p$ is selected to lie within the spectroscopic footprint, this action effectively regenerates a true dark region at a new location.

For example, consider the cell in Figure 6.11. It contains two dark regions. The compliment of these regions has a shape vaguely resembling a mushroom, as illustrated in Figure 6.12. One can relocate dark regions like these using the following procedure.

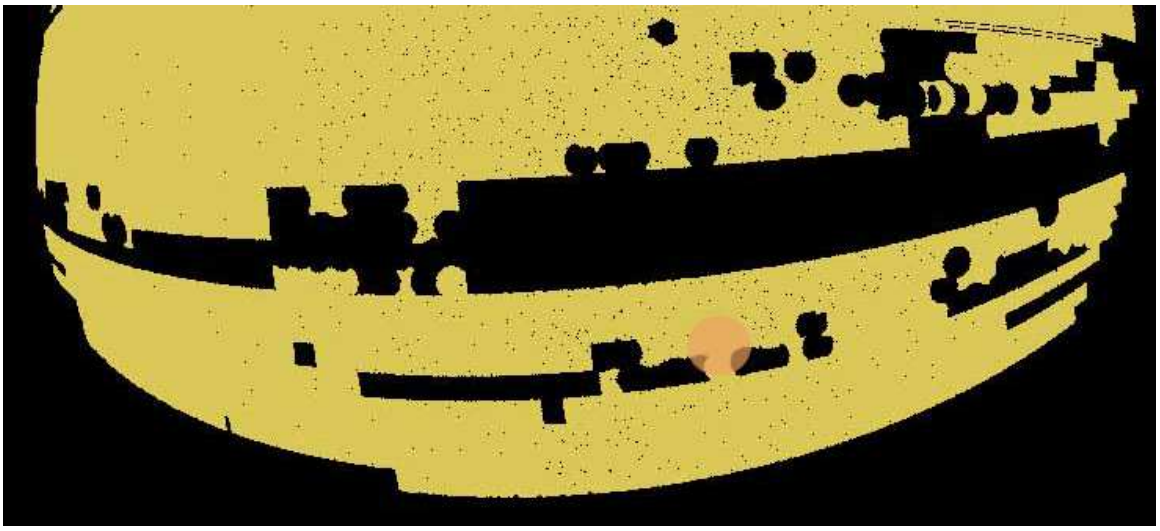

Figure 6.11: The circular projection of an R16 cell centered at $z = 0.113$ is shaded in sandy brown and superimposed upon the DR6 improved spectroscopic footprint as marked in yellow. This cell has two dark regions, one in the lower left and the other in the lower right.

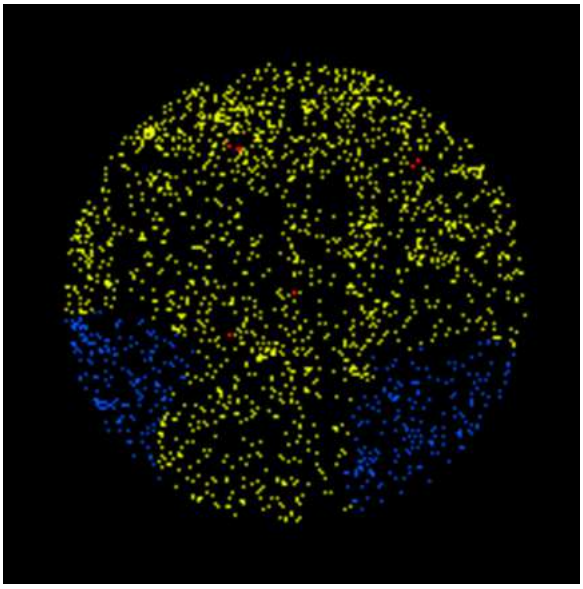

Figure 6.12: A view of all MGS targets that lie within the circular projection of the R16 cell at $z = 0.113$ from Figure 6.11. Targets within the spectroscopic footprint are represented by yellow pixels, while those outside are colored in blue or red. The seven red targets lie in the small areas in the compliment of TILEs.

First, from within the redshift range $[z_i, z_{i+1}]$, randomly select a cell that contains dark regions, i.e. where $\beta_{spec} < 0.99$. This will be referred to as a *template cell*. Then, select a random point $\boldsymbol{r}_p$ within the improved spectroscopic footprint and derive the rotation matrix $\mathbf{M}$ that recenters the template cell on that point. Multiply the spectroscopic footprint halfspace constraints by $\mathbf{M}$ and apply a random angular rotation about the cell's center to produce a shape like that pictured in Figure 6.13.

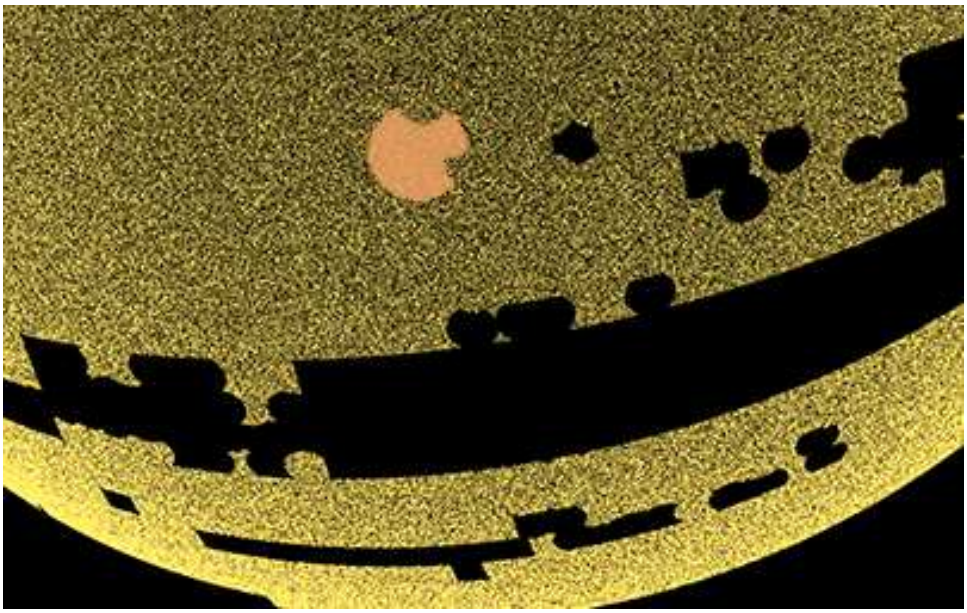

Figure 6.13: The cell (*sandy brown*) and its dark regions from Figure 6.11 are rotated into a new position within the spectroscopic footprint (*yellow*). Galaxies within the sandy brown layer retain their redshifts and become nearest neighbor candidates. Galaxies within the relocated dark regions will become *dark objects* and be stripped of their redshifts.

Next, identify the galaxies that fall within the relocated template cell's circular projection. Those that satisfy the rotated spectroscopic footprint constraint conditions retain their redshifts and are eligible to be nearest neighbors. The remainder become *dark objects* and are stripped of their redshifts. All cells within the redshift range $[z_i, z_{i+1}]$ whose circular

projections enclose at least one of the *dark objects* are identified and the nine counting methods are employed. This process continues until counts have been gathered for 10,000 cells in each redshift bin.

An astute observer might argue that requiring dark objects' nearest neighbors to lie within the rotated template cell is too restrictive. After all, there could be galaxies adjacent to the dark region but outside that cell. Disregarding what could *actually* be the nearest neighbors might cause the nearest neighbor and 2PCF smearing methods to appear weaker than they truly are.

We ignore this concern for two reasons. First, because dark regions tend to lie on the survey's edges, it is likely that there are no other galaxies on the side opposite the spectroscopic footprint. For example, consider the cell in Figure 6.11. The areas outside the cell, and adjacent to this cell's two dark regions *also* lie outside the spectroscopic footprint. No targets here can act as dark objects' nearest neighbors. In this case, the nearest neighbors are most likely to come from inside the template cell.

The second issue is practical. If we allowed nearest neighbors to come from outside the template cell, we would need to generate a larger list of candidate *galaxies*. To fairly represent the SDSS geometry, we would have to rotate all of the spectroscopic constraint conditions by $\mathbf{M}$ and compare them against a much larger set of targets. Nearest neighbor distances would have to be calculated. These additional computations would slow down the simulation process, leading to fewer cell counts overall.

While the majority of dark regions are caused by the survey's edges, some are due to

the small areas between TILEs (see the targets colored in red in Figure 6.12, for example). *Objects* that lie within them are classified as *dark objects*, even though their small number and close proximity to *galaxies* makes them more like interspersed objects.

While interspersed counting techniques would likely be more effective, we draw no distinction between dark objects based upon type of dark region they occupy. Doing so would require setting a limit beyond which regions are "small enough" to warrant a different treatment. At this juncture, quantifying such a limit would be arbitrary and without physical basis. The absolute number of dark objects in these regions is small, though, so treating them differently than those in the larger clumps is unlikely to significantly alter the results. At worst, failure to handle them separately would increase the error metric by some small amount. But since the goal of this analysis is to get a sense of which counting techniques are preferable to others, and since all counting techniques are subject to the same experimental conditions, we can expect the results to be generally applicable to the (larger) dark regions this section intends to study.

6.3.2.2 Results

Seven of the nine counting techniques could be applied to *dark objects* just as they were to *interspersed objects*. However, the "ignoring" and scaling methods, as well as determining the expected number of galaxies in each cell, require minor adjustments due to the introduction of the rotated constraint conditions. Each cell whose circular projection intersects a *dark object* has a precomputed spectroscopic completeness volume fraction

β_{spec}. Once the rotated constraint conditions are applied (i.e. those that generate the simulated dark regions), these cells adopt new completeness fractions $\beta' < \beta_{spec}$ that can be quantified using the usual Monte Carlo procedure.

Let the number of *galaxies* within a cell's volume after both sets of footprint conditions are applied be denoted by n_g. This serves as the final count for the "ignore" method. For the scaling method, the final number count n'_g equals the known *galaxy* count scaled by the relative increase in volume,

$$n'_g = n_g \frac{\beta_{spec}}{\beta'}. \tag{6.16}$$

Similarly, the expected number of galaxies in the cell once the dark region rotated constraints are implemented is

$$\langle n \rangle' = \langle n \rangle \frac{\beta_{spec}}{\beta'}. \tag{6.17}$$

The results of the dark region counting analysis for R7, R11, and R16 cells are presented in Figures 6.14, 6.15, and 6.16 respectively. A summary of the best counting methods at each redshift is supplied in Table 6.2. In brief, there is less uniformity in the preferred counting methods when dealing with dark regions than with interspersed regions. The counting results for dark regions are also more dependent upon the combination of redshift and cell size.

One characteristic shared by all cell sizes and measurement types (i.e. n, δ, δ^2) is that

	R7			R11			R16		
z	n	δ	δ^2	n	δ	δ^2	n	δ	δ^2
0.030	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.
0.045	scaling	scaling	scaling	D1sm.	D1sm.	D1sm.	D1sm.	D1sm.	SED
0.055	scaling	scaling	scaling	SEDsm.	SEDsm.	SED	SEDsm.	SEDsm.	SEDsm.
0.065	scaling	scaling	D1sm.	scaling	scaling	D1sm.	D1sm.	D1sm.	D1
0.075	scaling	scaling	2PCF	2PCF	2PCF	2PCF	D1sm.	D1sm.	D1
0.085	scaling	scaling	SEDsm.	scaling	SEDsm.	D1	SED	SED	D1
0.095	ignore	ignore	ignore	SEDsm.	SEDsm.	SEDsm.	SEDsm.	SEDsm.	D1
0.105	ignore	ignore	ignore	scaling	scaling	SEDsm.	SEDsm.	SEDsm.	SEDsm.
0.115	ignore	ignore	ignore	scaling	scaling	scaling	SEDsm.	D1sm.	D1
0.125	ignore	ignore	ignore	scaling	scaling	D1sm.	D1sm.	D1sm.	D1
0.135	ignore	ignore	ignore	D1sm.	D1sm.	D1	D1sm.	D1sm.	D1
0.145	ignore	ignore	ignore	D1sm.	D1sm.	D1sm.	D1	D1sm.	D1
0.155	ignore	ignore	ignore	D1sm.	D1sm.	D1	D1sm.	D1sm.	D1
0.165	ignore	ignore	ignore	D1sm.	D1sm.	D1	D1sm.	D1sm.	D1
0.175	ignore	ignore	ignore	ignore	ignore	ignore	D1sm.	D1sm.	D1
0.185	ignore	ignore	ignore	ignore	ignore	D1	D1sm.	D1sm.	D1
0.195	ignore	ignore	ignore	ignore	ignore	D1	D1sm.	D1sm.	D1
0.205	ignore	ignore	ignore	ignore	ignore	ignore	D1sm.	D1sm.	D1
0.215	ignore	ignore	ignore	ignore	ignore	ignore	ignore	ignore	D1sm.
0.225							ignore	ignore	D1
0.235							ignore	ignore	ignore
0.245							ignore	ignore	ignore
0.255							ignore	ignore	ignore
0.265							ignore	ignore	ignore
0.275							ignore	ignore	ignore
0.285							ignore	ignore	ignore
0.295							ignore	ignore	ignore

Table 6.2: A summary of the best methods to count dark objects for each cell size and measurement type as a function of redshift. Best methods are defined to be those with the lowest values of the error metrics $\epsilon^{(\,)}$.

the nearest neighbor method is the least successful. This is especially true at low redshifts, where large cellular projections — and consequently large dark regions — increase the average angular distance between dark objects and their nearest galaxy neighbors. The only time the nearest neighbor method offers any comparative advantage is for low-redshift R11 and R16 cells, where ignoring all objects outright sometimes turns out to be worse.

D1 photo-z smearing proves to be the best method for counting dark objects in cells between $z = 0.02$ and $z = 0.04$. Our presumption was that probabilistic methods would be most effective when counting large numbers of objects whose lines-of-sight intersect the same volume. Low-redshift cells certainly meet that description. Of the smearing methods tested, the probability distributions reported for D1 photometric redshifts would appear to be the most accurate in this redshift regime.

In the highest redshift cells, the optimal counting methods for dark objects share some properties with those for interspersed objects. In both cases, ignoring dark objects in the most distant cells (i.e. $z > [0.09, 0.17, 0.20]$ for R7/R11/R16) minimizes the error metric. As with interspersed objects, the redshifts at which ignoring objects is preferred increase with cell size.

The region/object types differ in that, for a given cell size, the transition point at which ignoring dark objects is preferable (to other counting methods) occurs at a lower redshift than for interspersed objects. This reveals a characteristic of dark regions. Because β_{spec} is fixed, the expected number of galaxies in dark regions decreases with distance. Therefore, Table 6.2 actually indicates the point at which the variance of dark-object-counts from other

methods exceeds the average number of dark objects *actually* in the cells.

For cells in low-to-intermediate redshifts (i.e. $0.04 < z < 0.09$ for R7, $0.04 < z < 0.17$ for R11, and $0.04 < z < 0.20$ for R16), the optimal counting methods differ considerably as a function of both a cell's size and redshift. Focusing for the moment on n and δ, we find that as with interspersed objects, scaling is best for R7 cells. For R16 cells, photometric redshifts of all varieties are preferred. While D1 photo-z smearing is the best option at most redshifts, SED photo-z smearing and discrete photometric redshifts are each optimal at select values of z.

For a higher resolution view of how other dark region counting methods compare to D1 photo-z smearing, reference Figures 6.17, 6.18, and 6.19. We call particular attention to the R11 results in Figure 6.18. Deciding which counting method is best at intermediate redshifts is something of a jumble. In this zoomed-in view, we see that the error metrics for different counting methods are often quite close to one another. Methods alternate between being preferable, to not, then back again.

These changes are small and rapid enough that it is worth asking how significant these preferences really are. For example, in the case of R11 cells, D1-smearing is favored at $z = 0.045$, SED-smearing is favored at $z = 0.55$, and scaling is favored at $z = 0.65$. Can we say with confidence SED-smearing is clearly best at $z = 0.55$, or are the methods so close in performance that any will suffice?

Figure 6.20 attempts to answer this question for the R11 case. Four pairs of counting methods are compared side-to-side in an attempt to quantify how much more effective

one method is over the other. The top subplot compares SED-smearing with D1-smearing across the cells' entire redshift range. At the lowest and highest redshifts, there is almost no comparison — D1-smearing dominates. At $z = 0.045$, D1-smearing outperforms the next-best option, SED-smearing, by slightly less than 1σ. This implies that D1-smearing is better than SED-smearing here about two-thirds of the time.

At the next transition point, $z = 0.065$, scaling outperforms the next closest option, SED-smearing. According to the second subplot of Figure 6.20, this result is also good to the 1σ level. Again, this is not definitive, but it is certainly significant. The performances of the two methods continue to follow each closely through $z = 0.105$. At these redshifts, however, the difference between their error metrics is mostly consistent with zero. (This conclusion might have been suspected from the fact that their curves nearly overlap in Figure 6.18.) In other words there is little, if any, benefit in preferring scaling or SED-smearing over the other when counting dark objects in cells at $0.08 < z < 0.105$.

At other redshifts, the differences between methods is more clear-cut. One situation in which the comparison is definitive is the transition between favoring D1-smearing at $z < 0.17$ and ignoring dark objects at $z > 0.17$. As the bottom subplot of Figure 6.20 illustrates, the distinction between the two is stark at all redshifts, save where the transition occurs.

Plots like Figure 6.20 for the R7 and R16 cases are omitted here for brevity, but we note that the significance properties are similar to those for R11. While the photo-z and smearing methods appear to deliver similar performance, the errors in $\Delta\epsilon^{(\,)}$ are usually small enough

that saying one method is preferred over another does carry statistical significance.

While the above comments hold for measures of n and δ, the error metrics $\Delta\epsilon^{(\delta^2)}$ for δ^2 display different characteristics. While the transitions to "ignoring" objects occur at similar redshifts, the preferred methods at smaller z tend to be more scattershot. In the majority of cases, the optimal δ^2 technique does not match those of n and δ. We are not unduly concerned with this finding, however, since statements of δ^2 optimality are less significant in general. Furthermore, our measurements of the power spectra derive from δ (even though δ^2 is arguably a more "natural" statistic), making it the quantity of higher priority.

Finally, we remind the reader that this analysis was performed on dark regions formed by cells for which $\beta_{spec} > 0.62$. The tests can be replicated for other values of β_{spec}, and the error metrics from Figures 6.14, 6.15, and 6.16 can be recalculated. This provides a way of quantifying the error that will result from one's choice of β_{spec} and may be used to reparameterize the minimum volume that one's cells must occupy within the spectroscopic footprint.

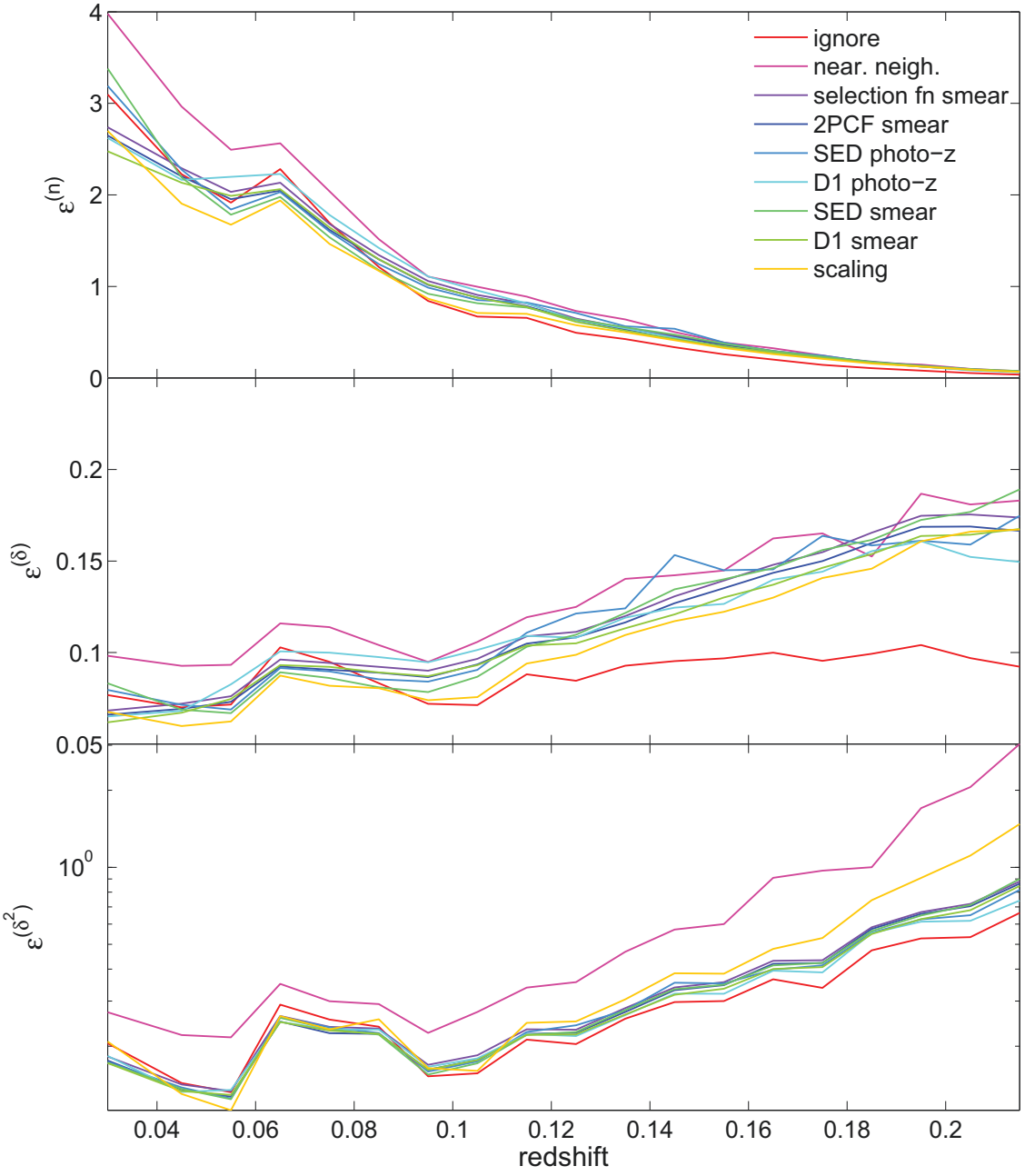

Figure 6.14: Dark region counting method results for R7 cells. Error metrics for number count $\epsilon^{(n)}$, overdensity $\epsilon^{(\delta)}$, and overdensity squared $\epsilon^{(\delta^2)}$ are presented on the vertical axis. Values are averaged over cells in redshift bins of width $\Delta z = 0.01$. Error bars are omitted for clarity, but are available elsewhere. Preferred counting methods have lower error metric values.

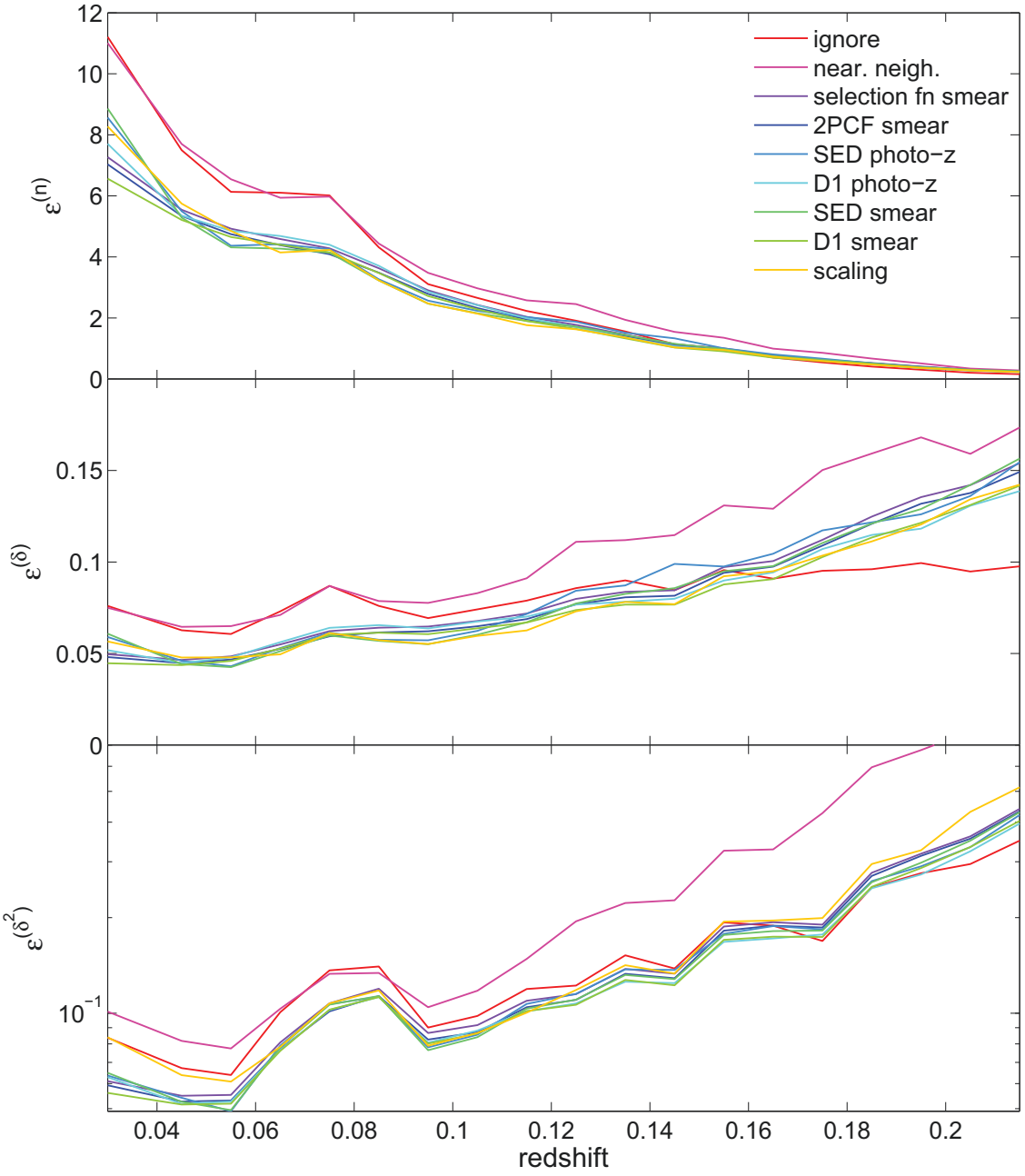

Figure 6.15: Same as Figure 6.14 but for R11 cells.

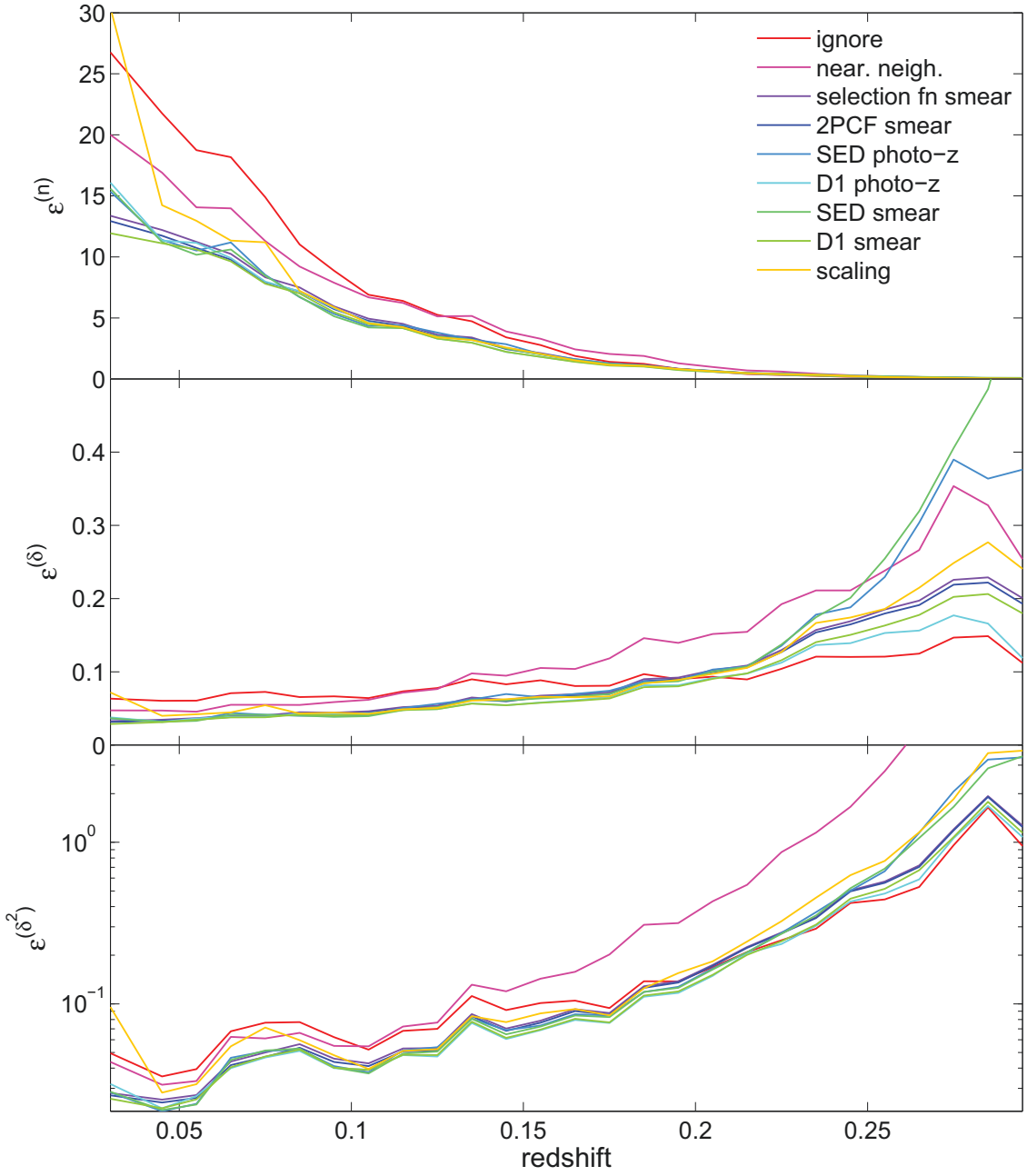

Figure 6.16: Same as Figure 6.14 but for R16 cells.

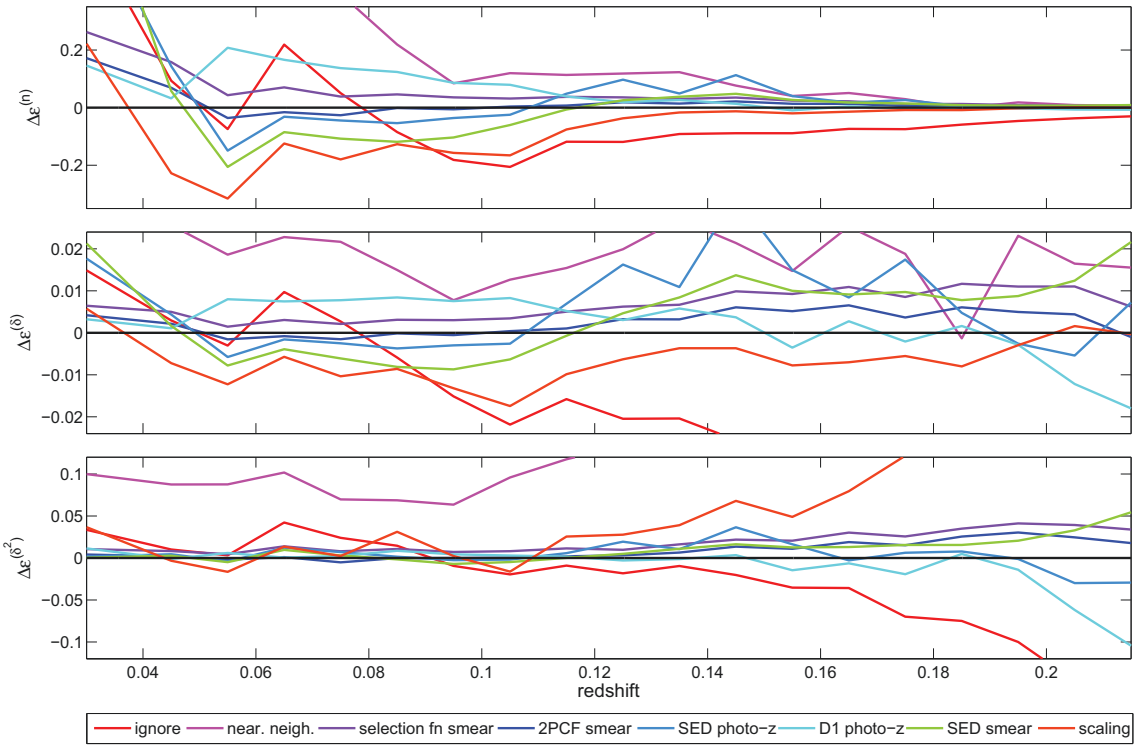

Figure 6.17: Comparison of error metrics for R7 dark regions relative to those for D1-smearing. This figure offers a more detailed view of the information presented in Figure 6.14. The vertical axis reports the difference in error metrics where $\Delta\epsilon^{(\,)}$ equals $\epsilon^{(\,)}$ for the methods indicated minus $\epsilon^{(\,)}$ for D1-smearing. At redshifts where $\Delta\epsilon^{(\,)} > 0$, D1-smearing is the better counting method. A counting technique with lower $\Delta\epsilon^{(\,)}$ at a given redshift is preferable to the alternative at that redshift.

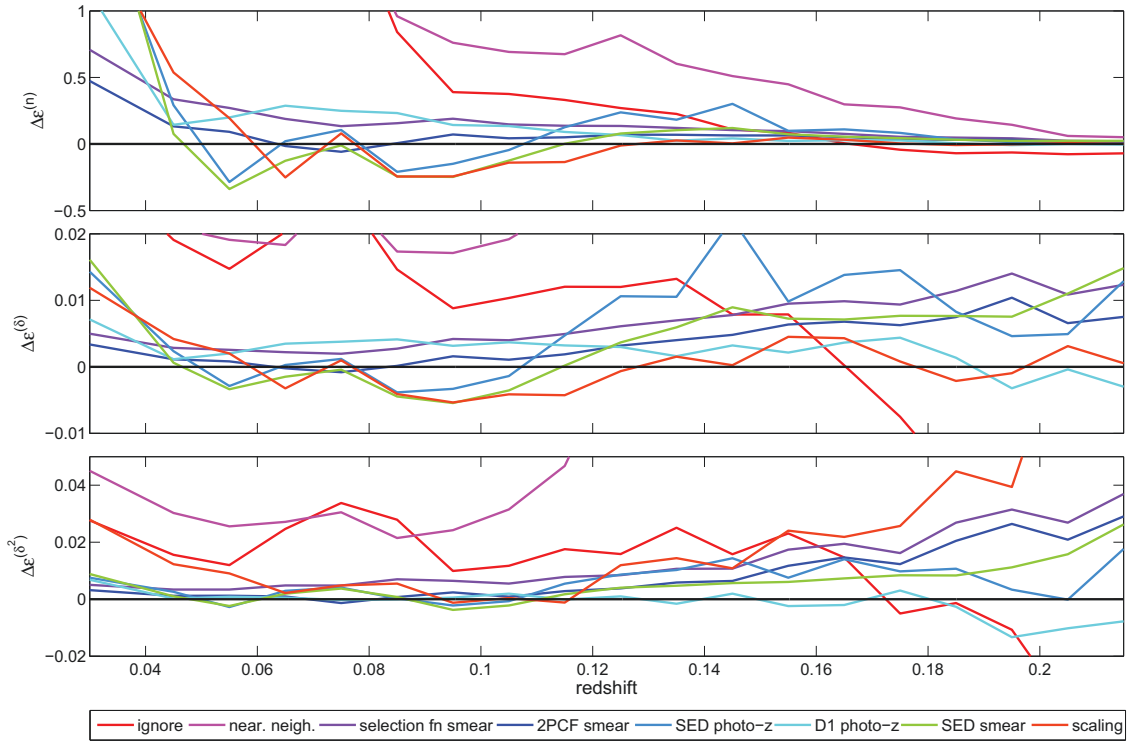

Figure 6.18: Same as Figure 6.17, but for R11 cells.

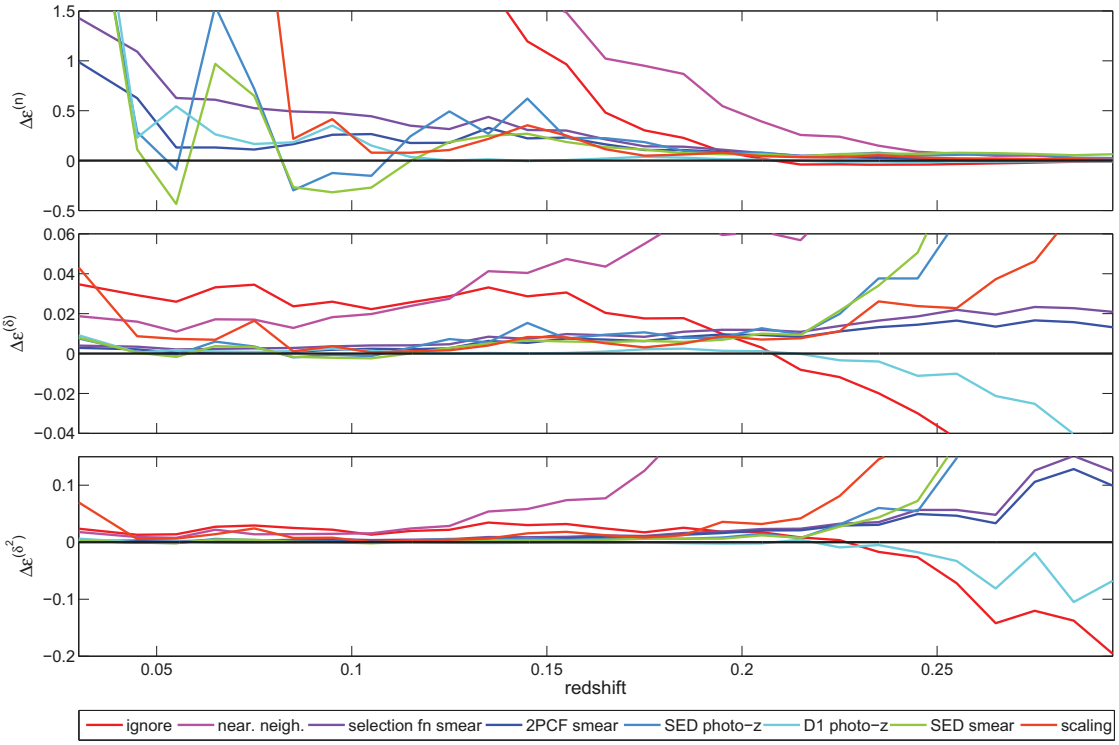

Figure 6.19: Same as Figure 6.17, but for R16 cells.

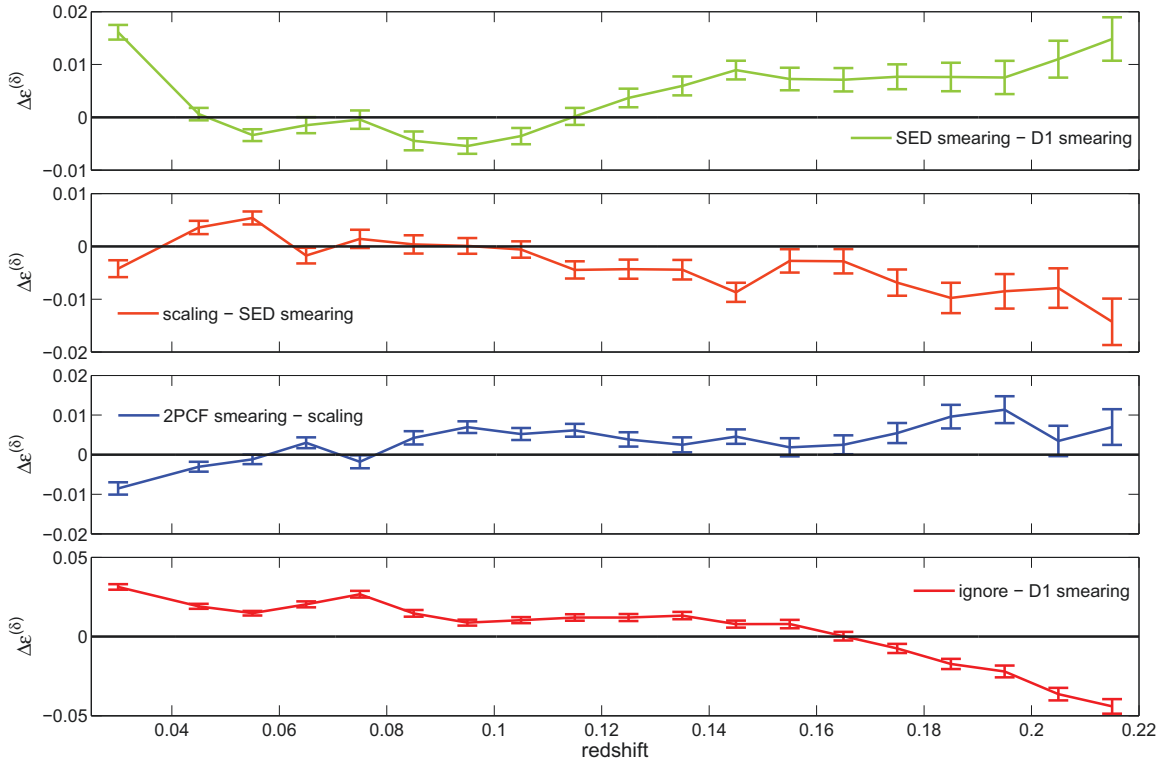

Figure 6.20: A comparison between select counting method pairs for dark objects in R11 cells. Each subplot reports the difference $\Delta\epsilon^{(\delta)}$ in the overdensity error metric $\epsilon^{(\delta)}$ between the two counting methods indicated. In the top subplot, for example, the curve traces $\epsilon^{(\delta)}$ for SED-smearing minus $\epsilon^{(\delta)}$ for D1-smearing. D1-smearing is preferred at redshifts for which $\Delta\epsilon^{(\delta)} > 0$. Errors bars are to 1σ.

6.3.3 External Regions

External regions are large, contiguous areas that contain no MGS galaxies. Unlike dark regions, which can occupy no more than a fraction $1 - \beta_{spec}$ of a cell's volume, external regions can be survey-sized. They are a natural byproduct of large photometric surveys prior to spectroscopic observation.

In this section, we describe how to simulate external regions from existing areas of the DR6 spectroscopic footprint. We apply our nine counting methods to these regions in an attempt to estimate the true number of galaxies in each cell. The results of this section can ultimately be used to inform strategies to handle survey areas where spectroscopic measurements have yet to occur.

6.3.3.1 Generating External Regions

We simulate external regions by carving away large areas near the edges of the DR6 improved spectroscopic footprint. The galaxies within them are subsequently stripped of their redshifts, becoming *external objects* in the process. The area that remains becomes the *trimmed spectroscopic footprint*, and the galaxies therein become candidates for the nearest neighbor and 2PCF smearing methods. As with interspersed and dark objects, all external objects are drawn from the pristine MGS galaxy sample.

The most straightforward way to trim the survey is by using the survey coordinate system $[\eta, \lambda]$ described in §2.2.1 and illustrated in Figure 2.6. These coordinates align with the direction of the STRIPEs and enable external regions to be carved away using simple

criteria.

Two sets of external regions that differ by size are created in the northern hemisphere. The first set, which we refer to as *External Region A*, is designed to capture about 4° worth of objects along the boundaries. MGS objects that satisfy any of the following eight conditions lie within External Region A:

- $\lambda < -20.1$ and $-14 < \eta < -7$
- $\lambda > -24.5$ and $-14 < \eta < -2.2$
- $-18 < \eta < -14$
- $\lambda < -13$ and $\eta < -29$
- $\lambda > -13$ and $\eta < -32$
- $\eta > 33$
- $\lambda < -57$
- $\lambda > 59$

The 125,306 galaxies that occupy External Region A are pictured in Figure 6.21. Color is used to represent the distance between each *external object* and its nearest neighbor. As seen in the figure, the nearest neighbor distances for many of these *objects* are so large that they likely lie beyond a spatial correlation radius that would put them at similar redshifts. For this reason, it is expected that the nearest neighbor method will perform poorly and that 2PCF smearing will offer little advantage over selection function smearing.

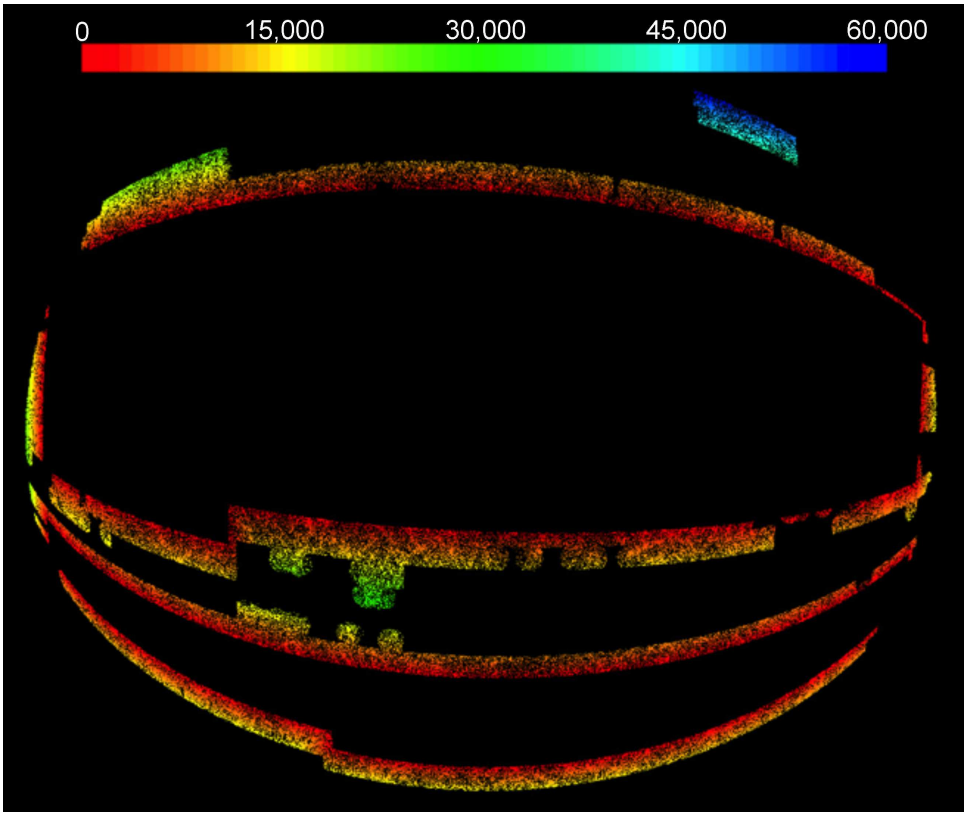

Figure 6.21: A view of External Region A. Each pixel marks an MGS galaxy that becomes an *external object* and is stripped of its redshift during the simulation process. Color represents the distance in arcseconds between each *external object* and its nearest neighbor within the trimmed spectroscopic footprint. The portions of External Region A have a characteristic length of about 4°.

Galaxies that satisfy any of the following eight criteria reside in *External Region B*. These limits are chosen so that the portions have characteristic lengths of about 2°. The 78,990 galaxies that lie within External Region B are pictured in Figure 6.22.

- $\lambda < -20.1$ and $-14 < \eta < -9$
- $\lambda > -24.5$ and $-14 < \eta < -4$
- $-18 < \eta < -14$
- $\lambda < -13$ and $\eta < -31$
- $\lambda > -13$ and $\eta < -34$
- $\eta > 34.3$
- $\lambda < -59.3$
- $\lambda > 60.7$

6.3.3.2 Results

We study external regions in an effort to understand how well our counting methods perform when *all* targets lack redshifts. Therefore, we prohibit any *galaxies* from within the trimmed spectroscopic footprint to be counted in cells. Because *galaxies* are disqualified from the counting analysis, the scaling method is undefined in external regions and consequently disregarded.

The exclusion of *galaxies* has the greatest impact on low-redshift cells and/or those with large angular projections. To avoid biasing our results, this change requires that we process our cell sample in a couple ways. To ensure uniformity, we only consider cells for which $\beta_{spec} > 0.99$. To account for the fact that cells that reach outside the external region

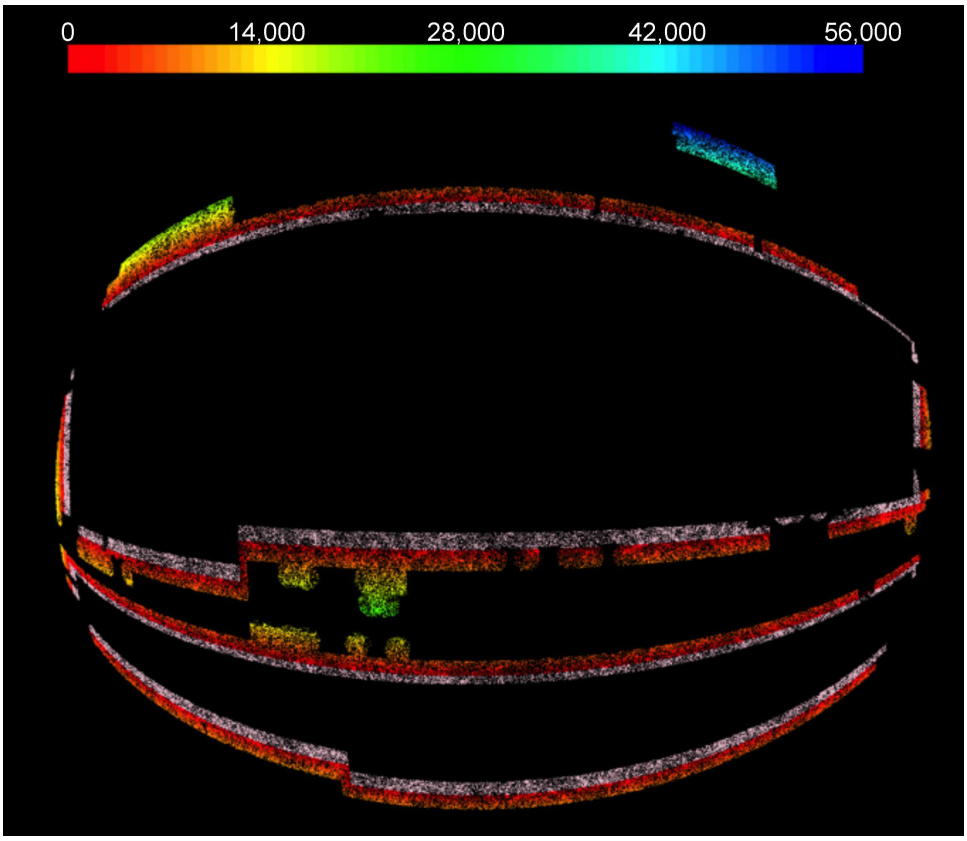

Figure 6.22: A view of External Region B. Each pixel on the red/blue spectrum marks an MGS galaxy that becomes an *external object* and is stripped of its redshift during the simulation process. Color represents the distance in arcseconds between each *external object* and its nearest neighbor within the trimmed spectroscopic footprint. The portions of External Region B have a characteristic length of about 2°. The galaxies colored in Mountbatten Pink lie within External Region A, but *not* External Region B. They are included for purposes of comparison.

will now contain "empty volumes", the volume each cell occupies in the external region *exclusively* is calculated using the usual Monte Carlo process. The number of galaxies expected therein is subsequently scaled downward to account for the volume reduction. Otherwise, tests proceed as normal.

The counting results for External Region A are presented Figures 6.23, 6.24, and 6.25. For all cell sizes and measurement types, the nearest neighbor method performs poorly

relative to the alternatives. Ignoring *objects* is also a very poor strategy, especially when approximating number count at low-z where the selection function is high. In select cases, the nearest neighbor and ignoring methods recover their efficacy at high redshifts where the true number of galaxies is frequently zero.

In estimating n, δ, and δ^2, the remaining methods perform similarly for all cell sizes. Figures 6.26, 6.27, and 6.28 help illustrate the relative differences between counting methods relative to D1-smearing. We find that SED photo-z's, D1 photo-z's, and SED-smearing offer roughly the same performance as D1-smearing, and slightly better than 2PCF/selection-function smearing.

In the R16 case, SED and D1-smearing perform exceptionally poorly when $z > 0.21$. This is due to large photo-z variances σ_z depositing too many partial counts at high redshifts. Ultimately, this reveals a problem with using a simple Gaussian model for the photo-z PDF's. Improvements might consist of a reduced high-z tail, better parameterized σ_z, or replacement of the Gaussian model with an individualized distribution for each target.

The results for External Region B are presented in Figures 6.29, 6.30, and 6.31. The conclusions are very similar to those of External Region A. None of the smearing or photo-z counting techniques emerge as clearly preferable to the others. At low redshifts, ignoring *objects* still performs poorly for all measures. The primary difference between the External Regions is that errors associated with the nearest neighbor method, while still significant, are smaller than in the External Region A case. This is a direct consequence of lowering the average angular distance between external objects and their nearest neighbors.

Tables 6.3 and 6.4 summarize the optimal counting methods for External Regions A and B respectively. As with interspersed and dark objects, ignoring external objects in high redshift cells still appears to be effective. D1-smearing performs well relative to other methods especially at $0.11 \leq z \leq 0.19$ although SED photo-z's, SED-smearing and 2PCF-smearing are also preferable in select circumstances. Photometric redshift smearing with Gaussian distributions fails at high redshifts.

In many cases, however, the preference for the optimal method over the others is not statistically significant. Figure 6.32 plots the differences in $\epsilon^{(\delta)}$ between D1-smearing and other counting methods for R11 cells in External Region A. On balanace, the 1σ uncertainties are many times larger than the differences themselves.

In summary, we conclude that no single photo-z or smearing method is significantly preferred over the others when attempting to count objects in external regions. The magnitudes of the error metrics $\epsilon^{(\,)}$ for external regions are approximately an order of magnitude larger than those for dark regions, though this comes with the caveat that cells containing dark regions are very likely to contain galaxies of known redshift as well as dark objects. Regardless, the external region error metrics are large enough to partially justify our constraint that cells must lie mostly within the spectroscopic footprint through the requirement $\beta_{spec} \geq 0.62$. Moreover, they support the conclusion derived from inspection of Figure 6.5 — none of the counting methods tested are capable of reproducing small scale structure.

	R7			R11			R16		
z	n	δ	δ^2	n	δ	δ^2	n	δ	δ^2
0.035	D1sm	D1sm	D1sm						
0.055	SEDsm	SED	SED	SED	SED	SED			
0.065	SED	SED	SED				D1sm	D1sm	D1sm
0.075	2PCF	2PCF	2PCF	SEDsm	SEDsm	D1sm			
0.085	SEDsm	SEDsm	SED	D1sm	D1sm	D1sm			
0.095	SEDsm	SEDsm	SED	SED	SED	SED	SED	SED	SED
0.105	SEDsm	SEDsm	SED	D1sm	D1sm	D1sm	SEDsm	SEDsm	D1sm
0.115	SEDsm	SEDsm	ignore	SEDsm	SEDsm	SED	D1sm	D1sm	D1sm
0.125	D1sm	D1sm	SEDsm	D1sm	D1sm	SEDsm	D1sm	D1sm	2PCF
0.135	D1sm	D1sm	ignore	D1sm	D1sm	SEDsm	D1sm	D1sm	SEDsm
0.145	D1sm	D1sm	ignore	D1sm	D1sm	D1sm	D1sm	D1sm	D1sm
0.155	D1sm	D1sm	ignore	D1sm	D1sm	D1sm	D1sm	D1sm	SF
0.165	D1sm	D1sm	ignore	D1sm	D1sm	SEDsm	D1sm	D1sm	SED
0.175	D1sm	D1sm	ignore	D1sm	D1sm	SEDsm	D1sm	D1sm	D1sm
0.185	ignore	ignore	ignore	D1sm	D1sm	SED	D1sm	D1sm	D1sm
0.195	ignore	ignore	ignore	D1sm	D1sm	ignore	D1sm	D1sm	D1sm
0.205	ignore	ignore	ignore	SED	SED	ignore	SED	SED	2PCF
0.215	ignore	ignore	ignore	ignore	ignore	ignore	SED	SED	SED
0.225							SED	SED	SED
0.235							2PCF	2PCF	ignore
0.245							2PCF	2PCF	ignore
0.255							2PCF	2PCF	ignore
0.265							ignore	ignore	ignore
0.275							ignore	ignore	ignore
0.285							ignore	ignore	ignore
0.295							ignore	ignore	ignore

Table 6.3: A summary of the best methods to count external objects in External Region A for each cell size and measurement type as a function of redshift. Best methods are defined to be those with the lowest values of the error metrics $\epsilon^{(\,)}$. Redshift bins containing too few cells to generate meaningful statistics are grouped together.

	R7			R11			R16		
z	n	δ	δ^2	n	δ	δ^2	n	δ	δ^2
0.040	2PCF	2PCF	SED						
0.065	D1sm	D1sm	SEDsm	SF	SF	SED			
0.075	2PCF	2PCF	SED						
0.085	SEDsm	SEDsm	ignore				SED	SED	SED
0.095	SEDsm	SEDsm	SEDsm	SED	SED	SED			
0.105	SEDsm	SEDsm	ignore	D1sm	D1sm	D1sm			
0.115	SEDsm	SEDsm	ignore	D1sm	D1sm	SED			
0.125	D1sm	D1sm	D1sm	D1sm	D1sm	D1sm	SEDsm	SEDsm	SF
0.135	D1sm	D1sm	ignore	SED	SED	SED			
0.145	D1sm	D1sm	ignore	D1sm	D1sm	2PCF	D1sm	D1sm	SED
0.155	D1sm	D1sm	ignore	D1sm	D1sm	SED	D1sm	D1sm	SEDsm
0.165	D1sm	D1sm	ignore	D1sm	D1sm	D1sm	2PCF	2PCF	SED
0.175	D1sm	D1sm	ignore	D1sm	D1sm	SEDsm	D1sm	D1sm	2PCF
0.185	ignore	ignore	ignore	D1sm	D1sm	SED	D1sm	D1sm	D1sm
0.195	ignore	ignore	ignore	D1sm	D1sm	ignore	2PCF	2PCF	SF
0.205	ignore	ignore	ignore	SED	SED	ignore	2PCF	SED	SF
0.215	ignore	ignore	ignore	ignore	2PCF	ignore	SED	SED	ignore
0.225							2PCF	ignore	ignore
0.235							ignore	ignore	ignore
0.245							2PCF	2PCF	ignore
0.255							ignore	ignore	ignore
0.265							ignore	ignore	ignore
0.275							ignore	ignore	ignore
0.285							ignore	ignore	ignore
0.295							ignore	ignore	ignore

Table 6.4: Same as Table 6.3 but for External Region B.

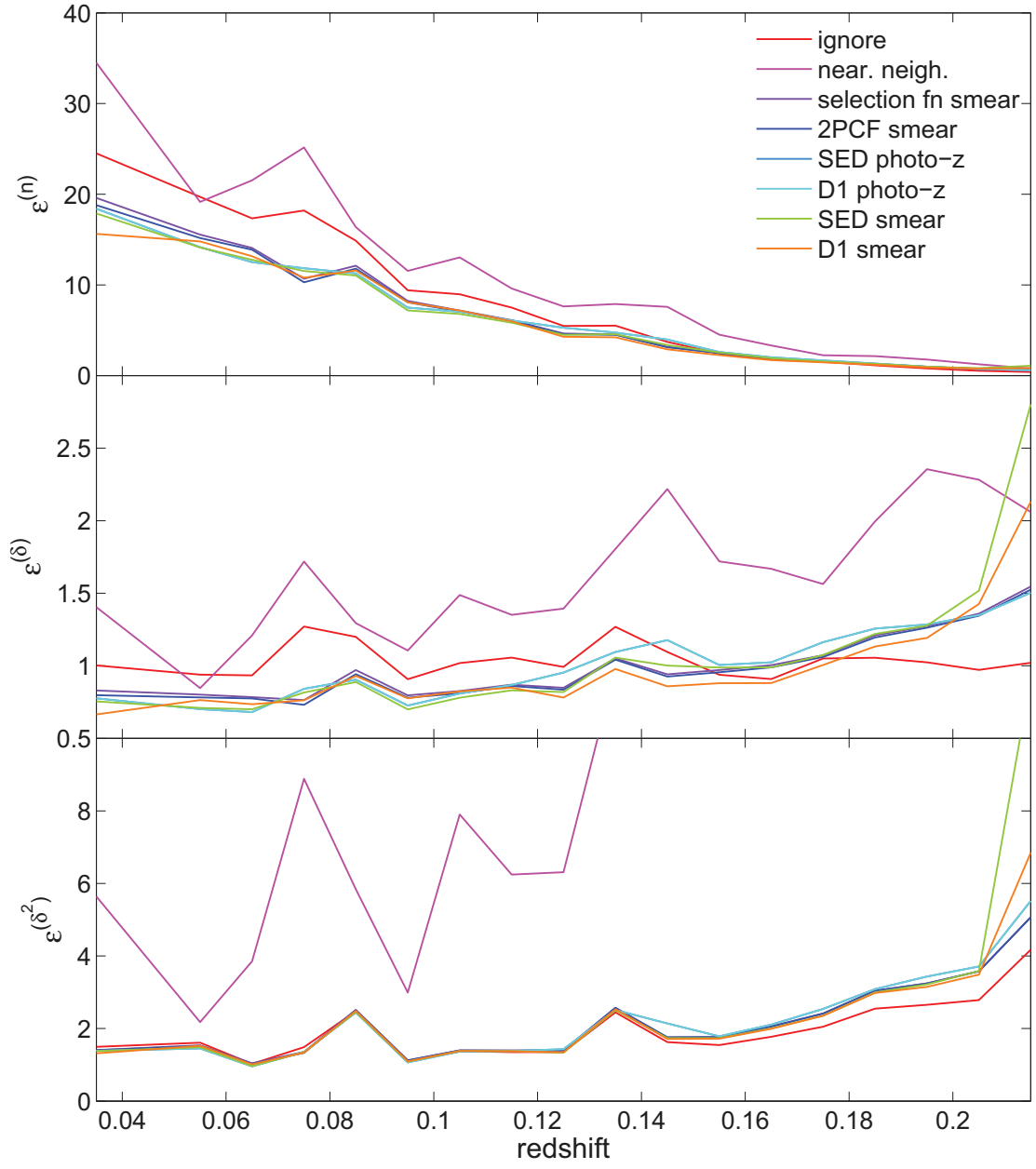

Figure 6.23: External Region A counting method results for R7 cells. Error metrics for number count $\epsilon^{(n)}$, overdensity $\epsilon^{(\delta)}$, and overdensity squared $\epsilon^{(\delta^2)}$ are presented on the vertical axis. Values are averaged over cells in redshift bins of width $z = 0.01$. Error bars are omitted for clarity, but are available in text files online. Preferred counting methods have lower error metric values.

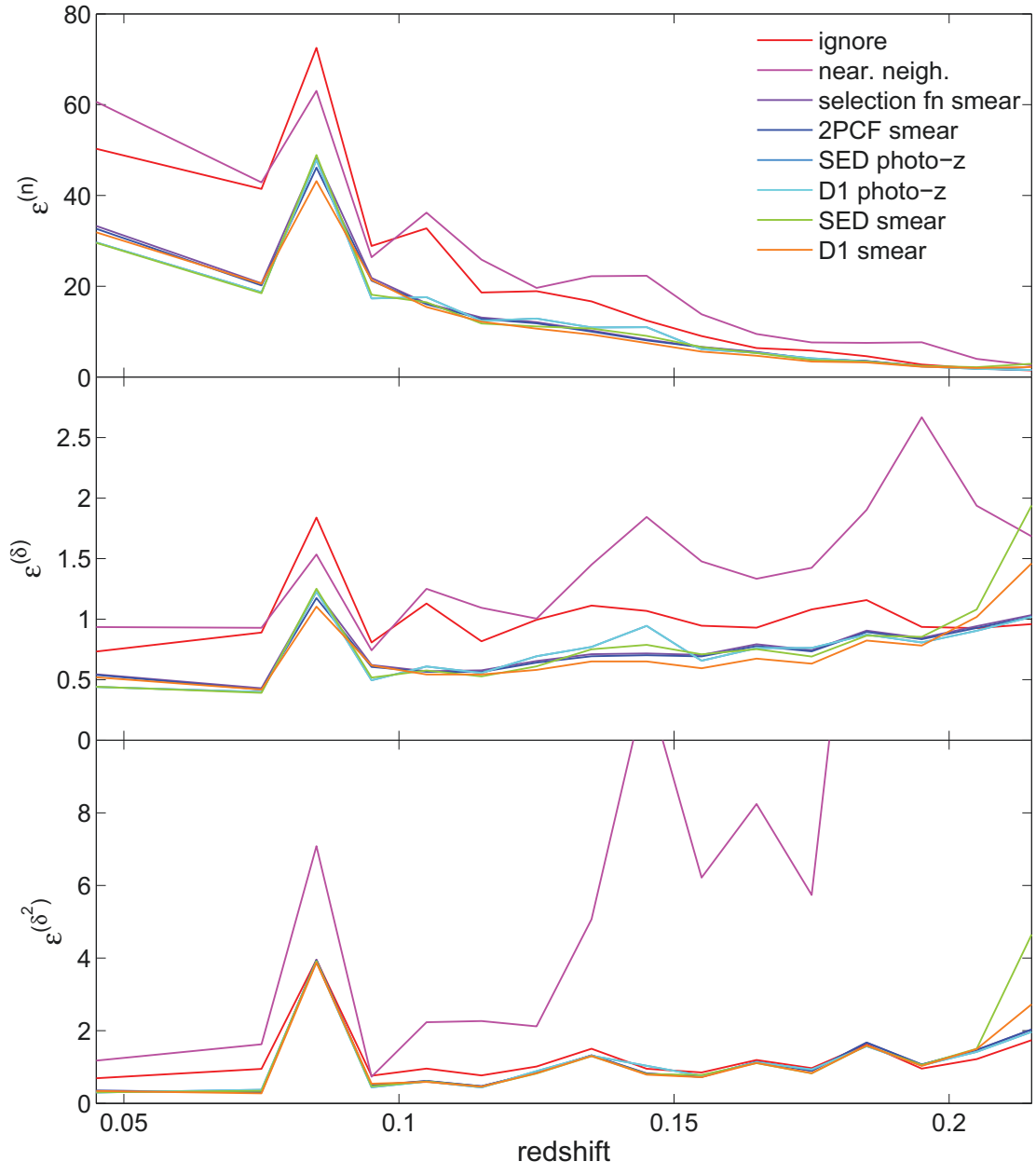

Figure 6.24: Same as Figure 6.23 but for R11 cells.

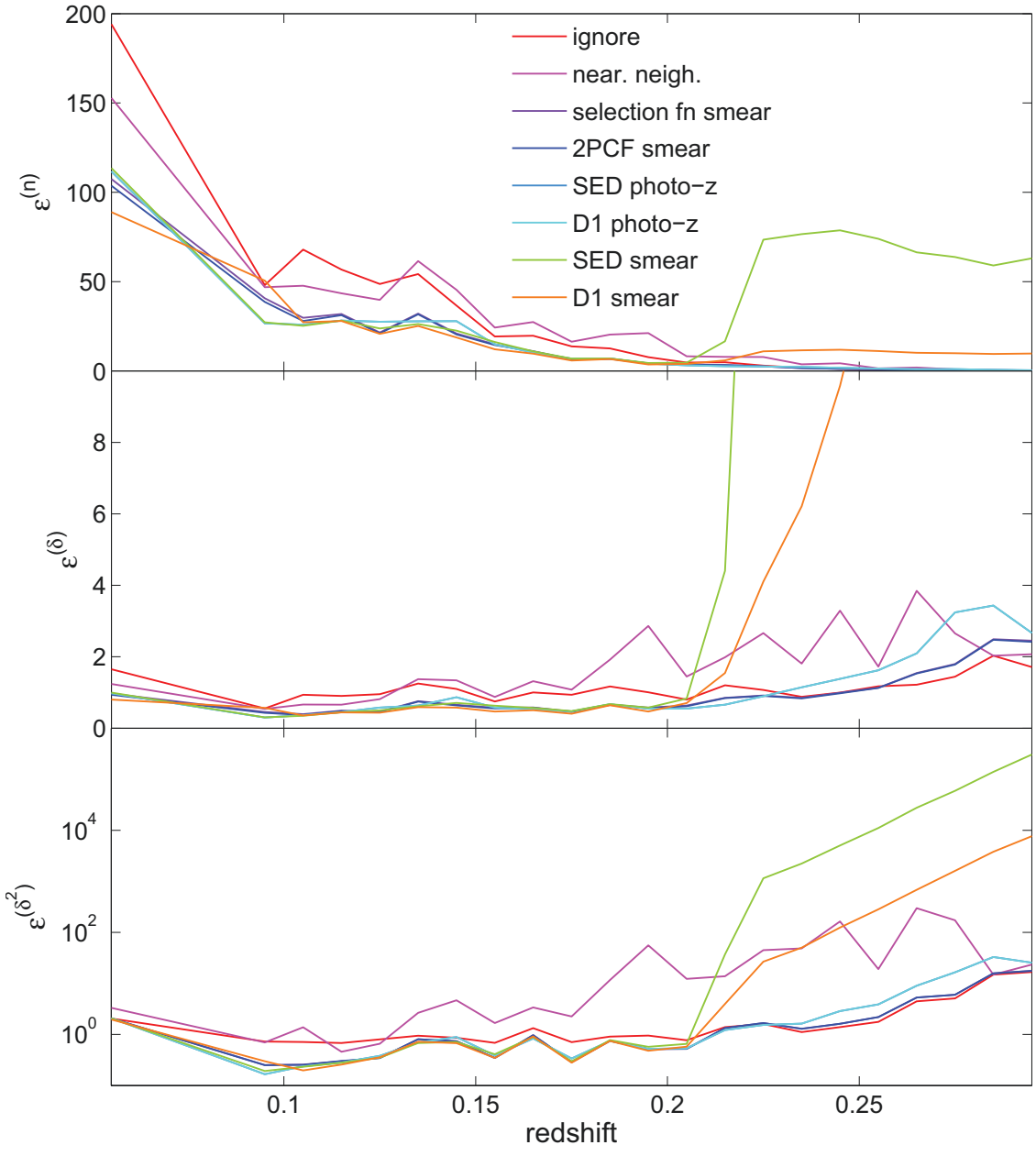

Figure 6.25: Same as Figure 6.23 but for R16 cells.

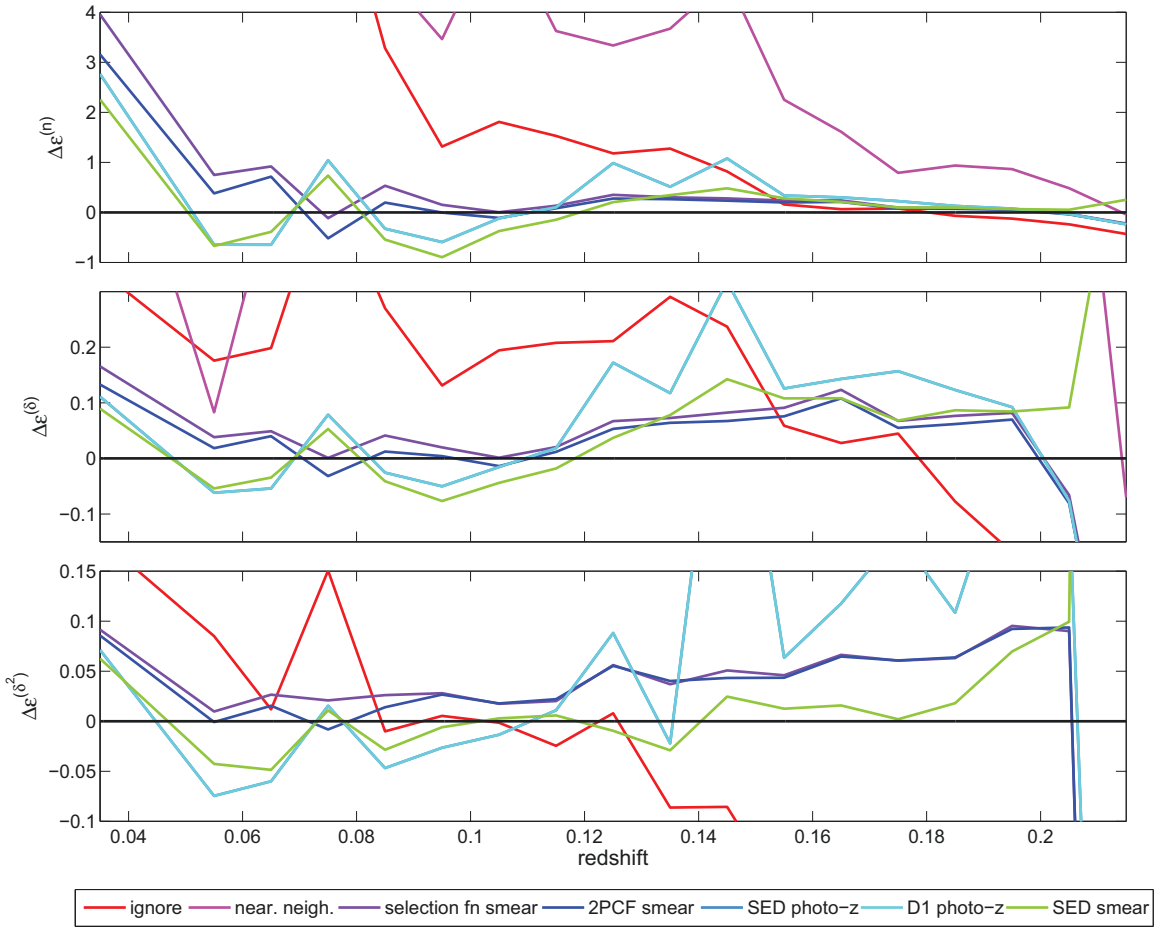

Figure 6.26: Comparison of error metrics for R7 External Region A relative to those for D1-smearing. This figure offers a more detailed view of the information presented in Figure 6.23. The vertical axis reports the difference in error metrics where $\Delta\epsilon^{(\,)}$ equals $\epsilon^{(\,)}$ for the methods indicated minus $\epsilon^{(\,)}$ for D1-smearing. At redshifts where $\Delta\epsilon^{(\,)} > 0$, D1-smearing is the better counting method. A counting technique with lower $\Delta\epsilon^{(\,)}$ at a given redshift is preferable to the alternative at that redshift.

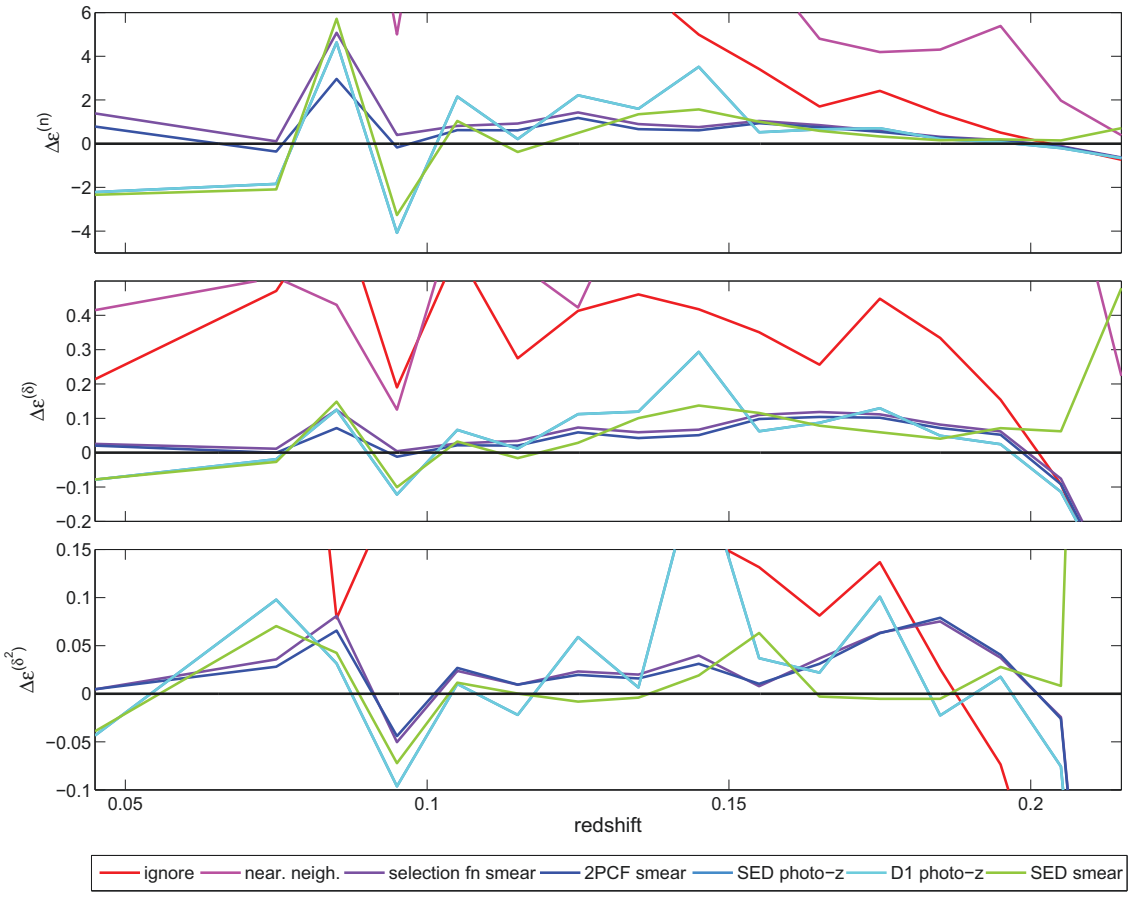

Figure 6.27: Same as Figure 6.26 but for R11 cells.

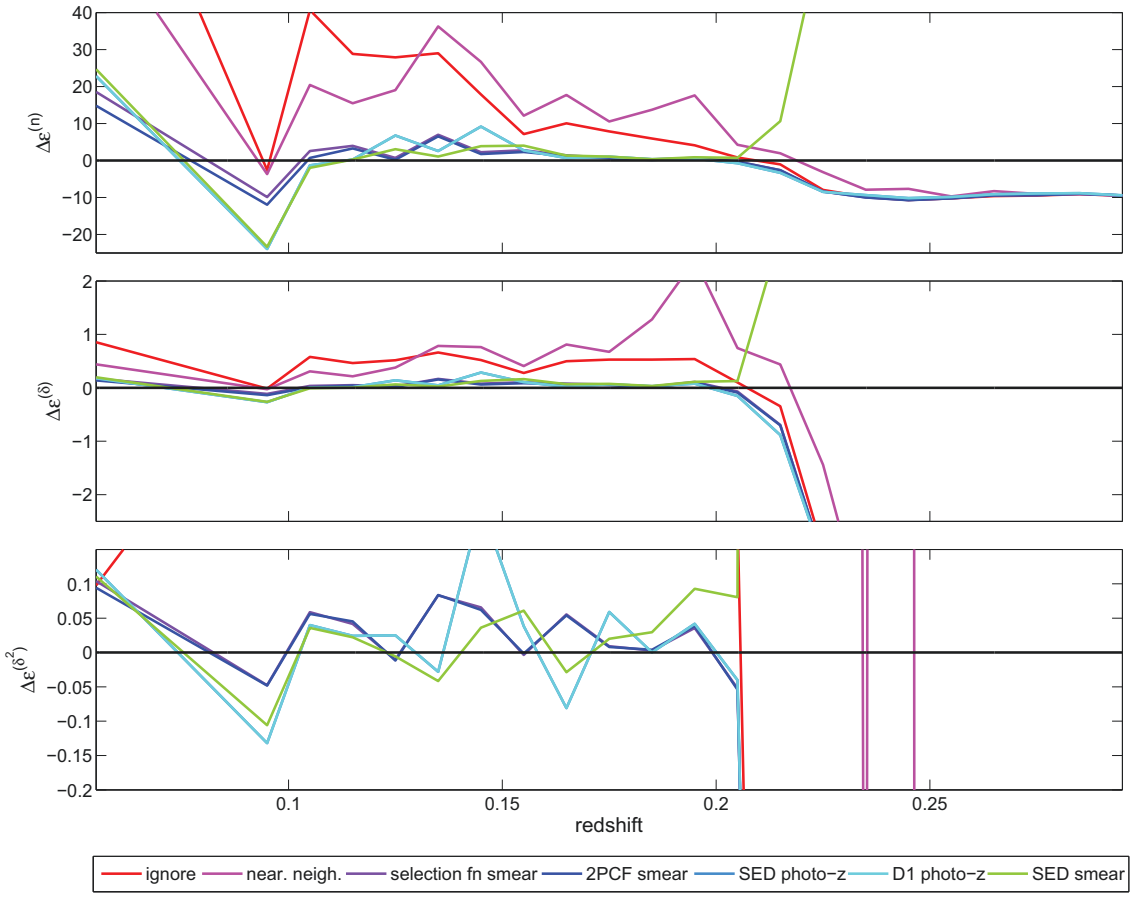

Figure 6.28: Same as Figure 6.26 but for R16 cells.

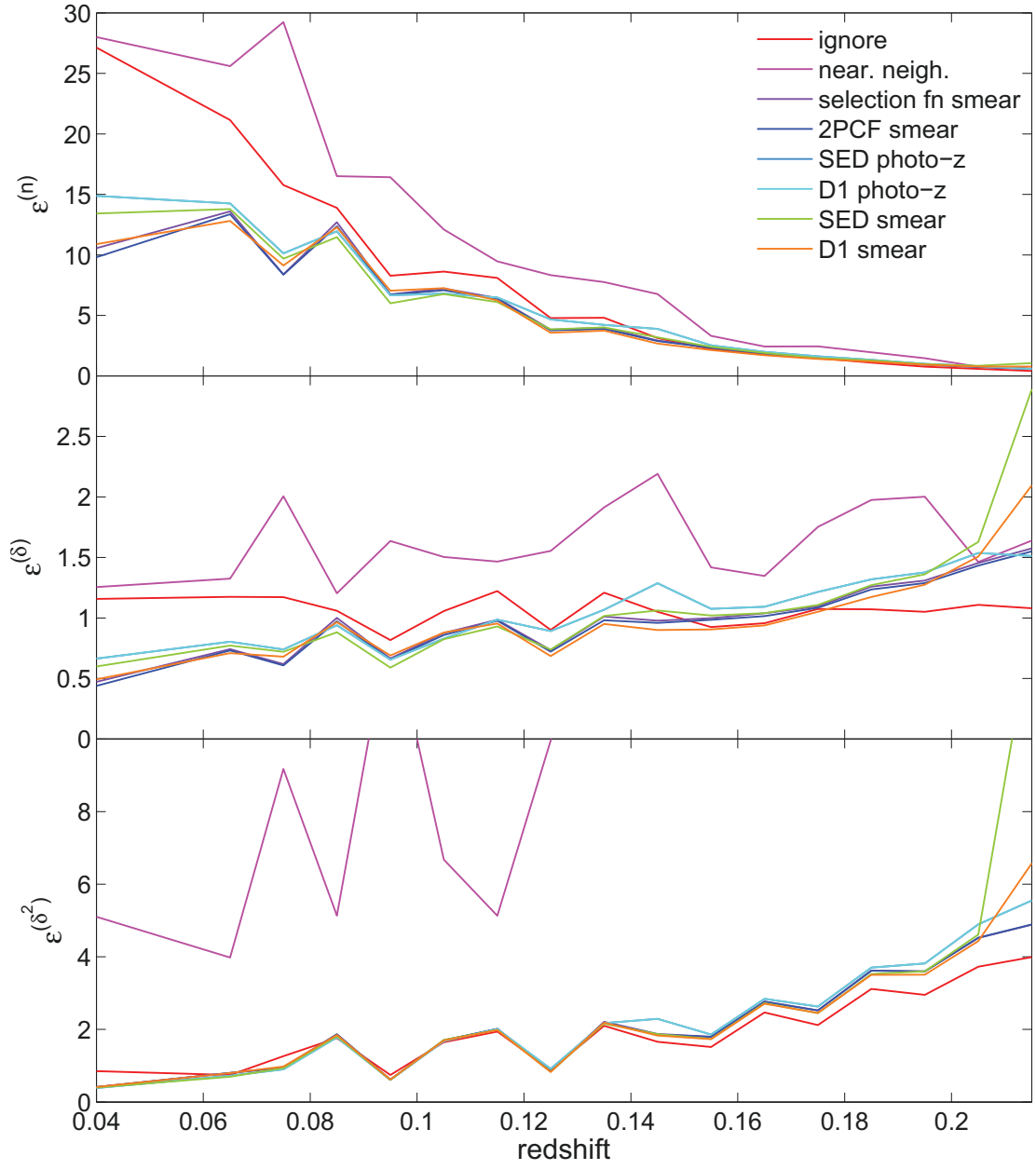

Figure 6.29: External Region B counting method results for R7 cells. Error metrics for number count $\epsilon^{(n)}$, overdensity $\epsilon^{(\delta)}$, and overdensity squared $\epsilon^{(\delta^2)}$ are presented on the vertical axis. Values are averaged over cells in redshift bins of width $z = 0.01$. Error bars are omitted for clarity, but are available in text files online. Preferred counting methods have lower error metric values.

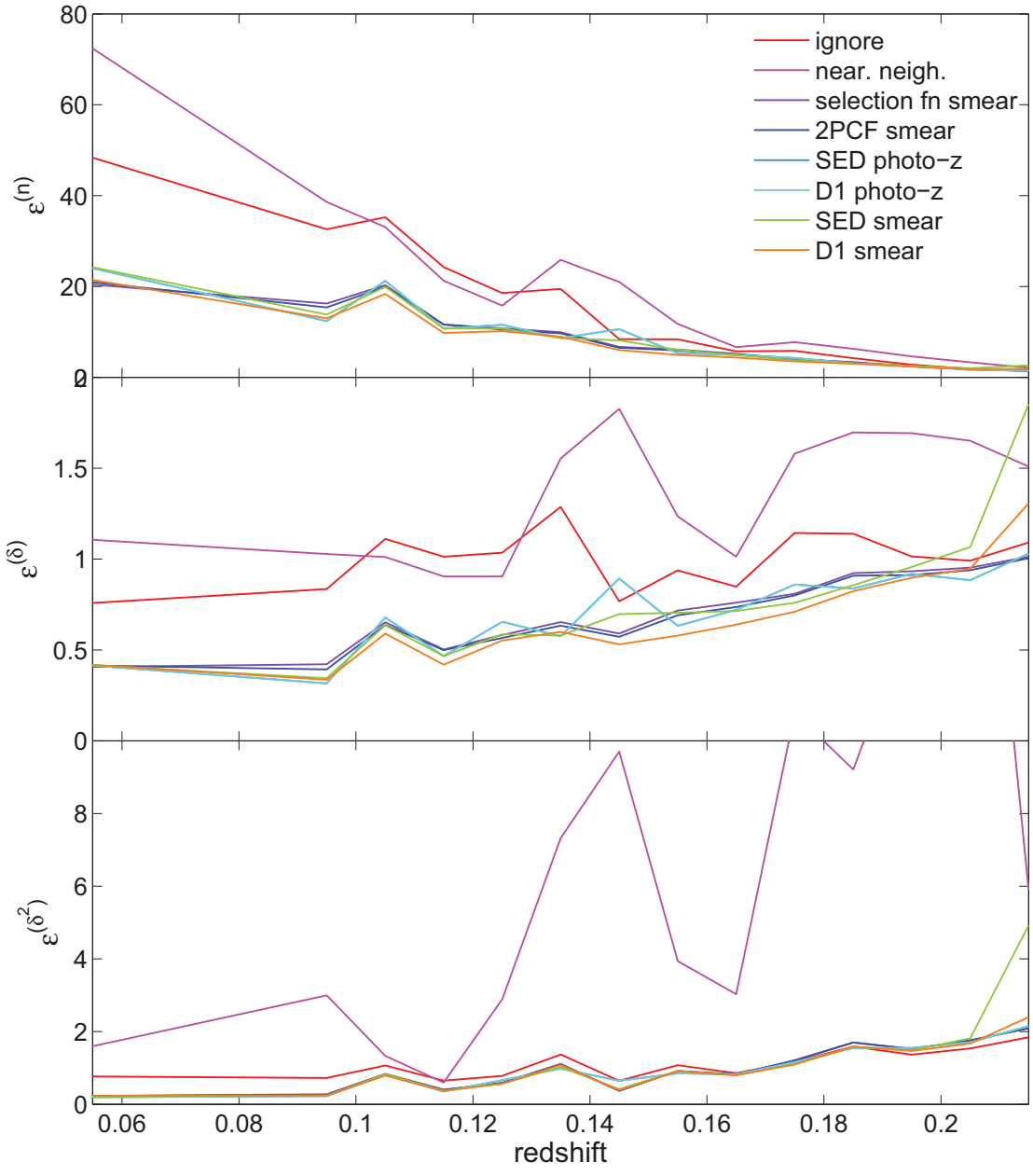

Figure 6.30: Same as Figure 6.29 but for R11 cells.

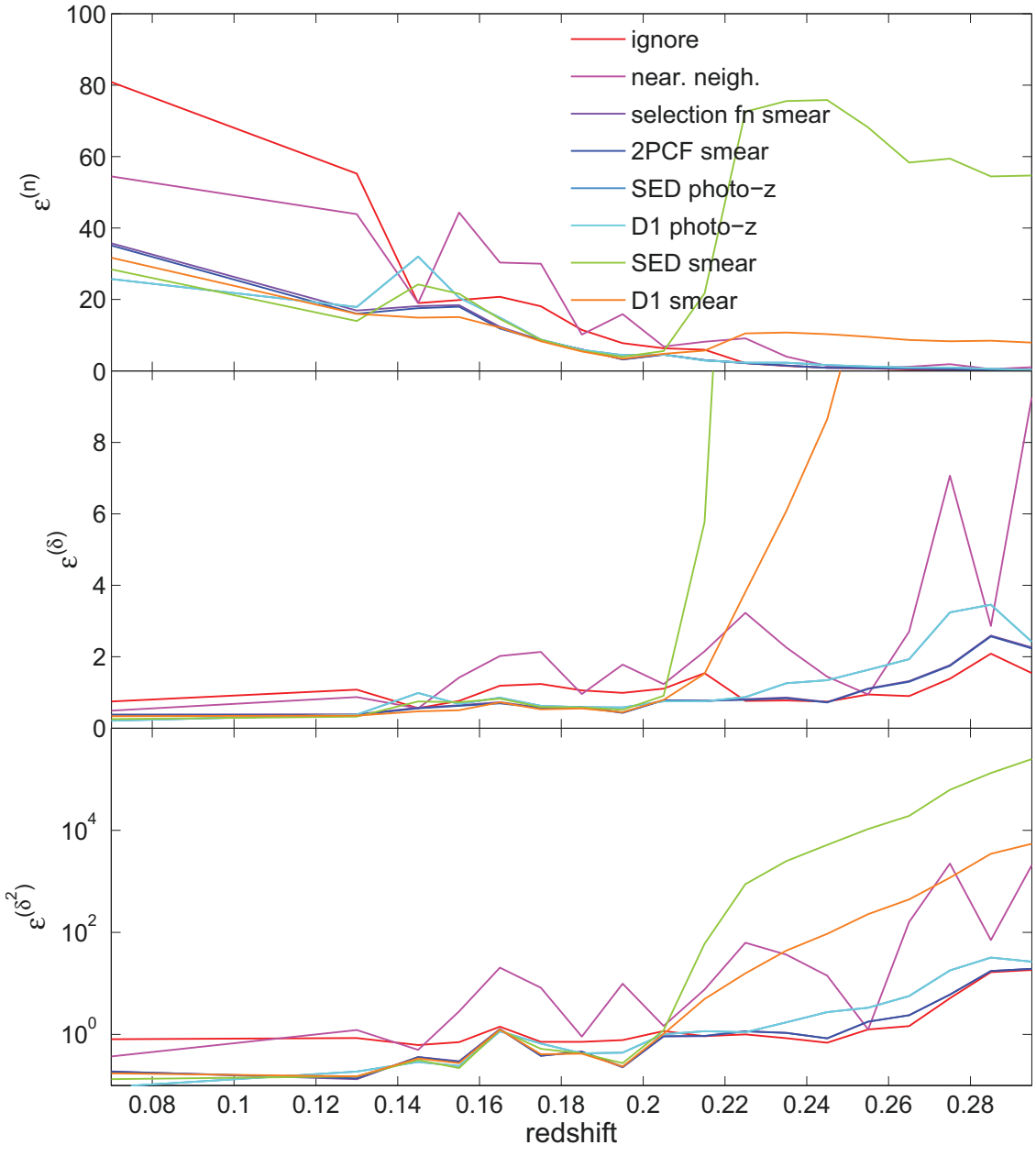

Figure 6.31: Same as Figure 6.29 but for R16 cells.

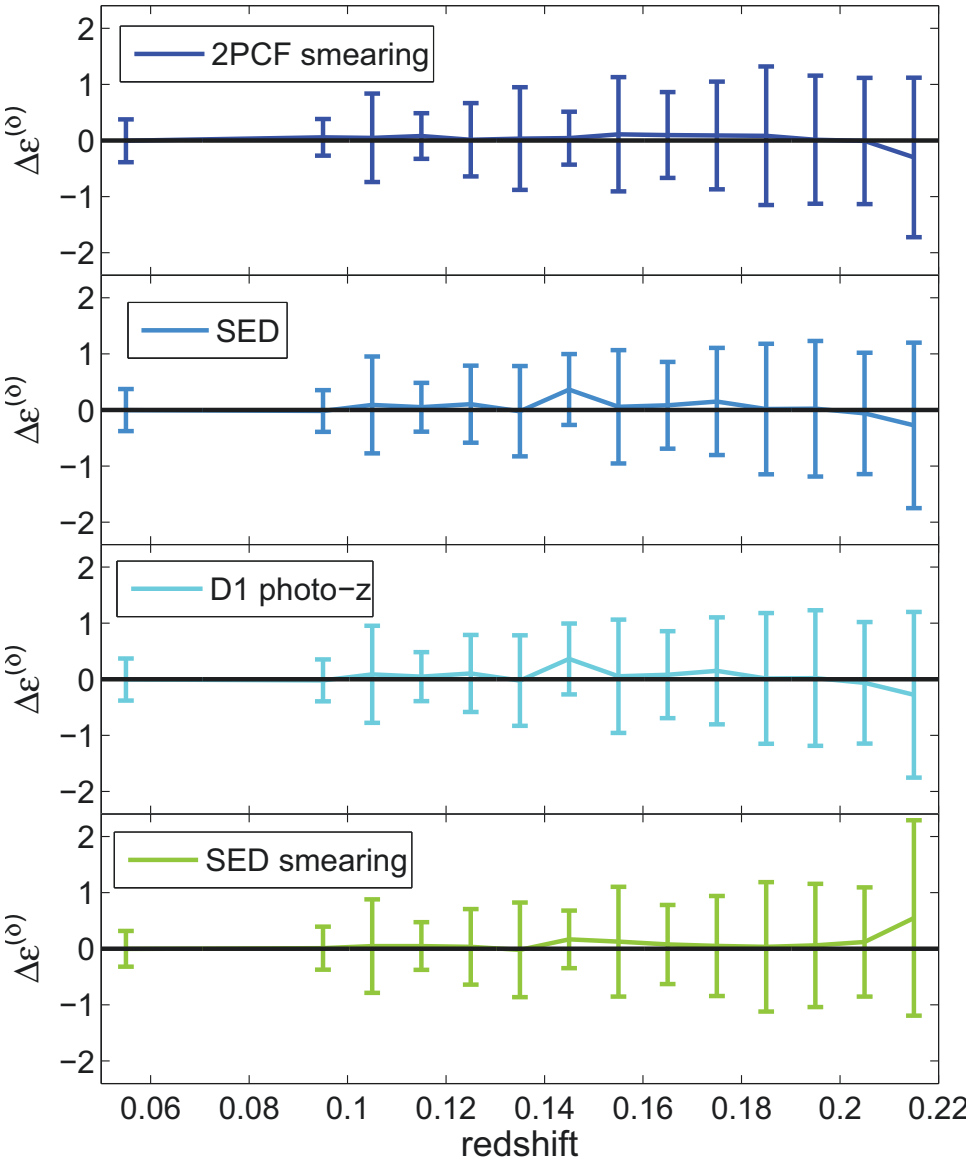

Figure 6.32: Comparison between $\epsilon^{(\delta)}$ for D1-smearing and other select methods for R11 cells in External Region A. The vertical axis reports the difference in error metrics where $\Delta\epsilon^{(\delta)}$ equals $\epsilon^{(\delta)}$ for the methods indicated minus $\epsilon^{(\delta)}$ for D1-smearing. At redshifts where $\Delta\epsilon^{(\delta)} > 0$, D1-smearing is the better counting method. The magnitudes of the 1σ spreads in the differences of the means are large enough to render statements of D1-smearing's optimality over other methods to be of low significance.

6.3.4 Additional Comparisons

In this section, we consider two additional comparisons of counting techniques — photometric redshifts vs. photo-z smearing, and selection function smearing vs. 2PCF smearing. We only consider results from the interspersed and dark regions analyses since most of the external regions work was inconclusive.

We begin by examining the differences between photo-z's and photo-z smearing. For each measurement type and at each redshift, we identify the photometric redshift (i.e. SED or D1) that produced the minimum value of the error metric. We do the same for photometric smearing to arrive at a new "best photo-z" and "best smearing photo-z" statistic. This removes any distinction between the type of photometric redshift used and, instead, compares discrete versus probabilistic smearing with photo-z's on their own.

In Figure 6.33 we plot the differences in error metrics between the "best" discrete and smeared photometric redshifts for interspersed and dark regions.[3] Several conclusions emerge. First, smearing tends to fail in regions where $\langle n \rangle$ is low. These include R16 cells at $z \gtrsim 0.22$ and R7 cells at $z \gtrsim 0.17$.[4] At these higher redshifts, excess probability in the photo-z PDF's drives up measures of n. This leads to the conclusions that at high-z, all else being equal, discrete photometric redshifts are preferable to smeared photometric redshifts.

The second conclusion is that for low and intermediate redshift cells, probabilistic

[3]Error bars for interspersed regions are smaller than those for dark regions. This is partly by virtue of the fact that interspersed counts use 50,000 cells per redshift bin while dark regions only use 10,000.

[4]Measures of δ and δ^2 are very sensitive to the expected number of galaxies, a value which increases with redshift. This explains the relatively large variances at high z.

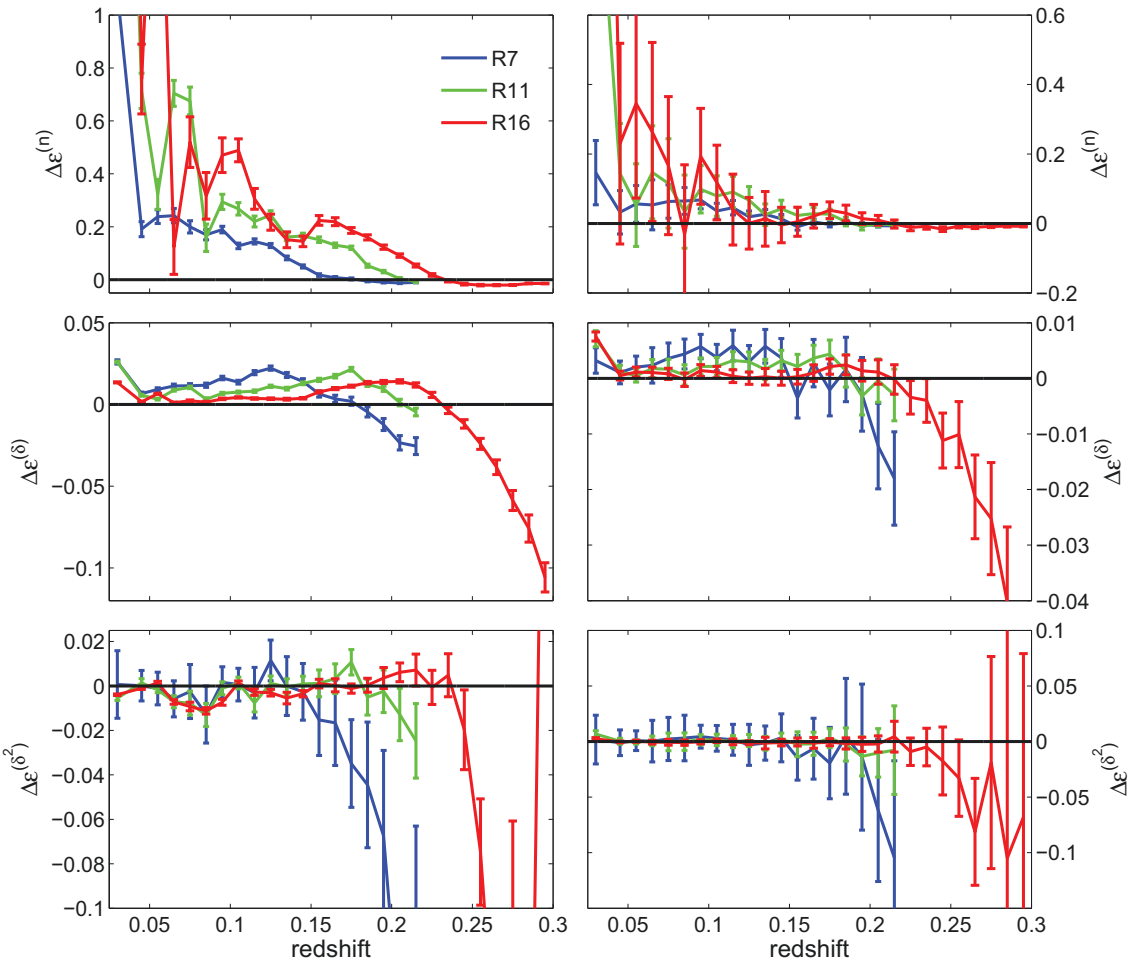

Figure 6.33: A comparison between photo-z's and photometric redshift smearing as it relates to approximating counting statistics. For each cell size and redshift bin, the optimal photo-z (SED or D1) and smeared photo-z counting methods are determined. The errors $\epsilon^{(\delta)}$ for the best photometric redshift smearing techniques are subtracted from the errors $\epsilon^{(\delta)}$ for the best photo-z counting methods to produce a comparison statistic $\Delta\epsilon^{(\delta)}$. When $\Delta\epsilon^{(\delta)} > 0$, photometric redshift smearing is outperforming the use of photometric redshifts alone. Similar comparisons for $\Delta\epsilon^{(n)}$ and $\Delta\epsilon^{(\delta^2)}$ are made and presented in rows. The left and right columns contain the results for interspersed and dark regions respectively.

photo-z's are generally better for measuring n and δ, while discrete photo-z's preferable for δ^2 (though this latter preference is of limited statistical significance). This conclusion holds for both interspersed and dark regions. This is an anticipated result since the regions differ

only in the way objects are clustered within the cells — a distinction that has no impact on the efficacy of photometric redshift methods.

Finally, we conclude that photo-z smearing methods offer relatively better performance in measuring n when $\langle n \rangle$ is large. Recall the assumption underlying probabilistic smearing — when applied in aggregate the sum of partial counts should approach the true count. This condition is best realized when the number of targets in each cell is high, as is the case with the low-redshift R16 cells in Figure 6.33.

For the second comparison, we explored the differences between probabilistic smearing using the selection function versus the 2PCF. The results are plotted in Figure 6.34. To summarize the figure in one statement: *2PCF-smearing is always preferable to selection function smearing.* The information "boost" provided to the selection function by including spatial correlation information from an object's nearest neighbor is always beneficial in both interspersed and dark regions.

This benefit is larger and more significant for interspersed objects than dark objects. Because the average nearest neighbor distance is smaller for interspersed objects, the boost provided by the correlation function is more valuable here than in dark regions. We also note that the boost is better for larger cells when approximating n, and for smaller cells when approximating δ and δ^2.

There are many other comparisons that could be made between counting methods as a function of redshift, region variety, and measurement type. For example, one might wish to know which discrete counting method is best at each redshift and how these compare

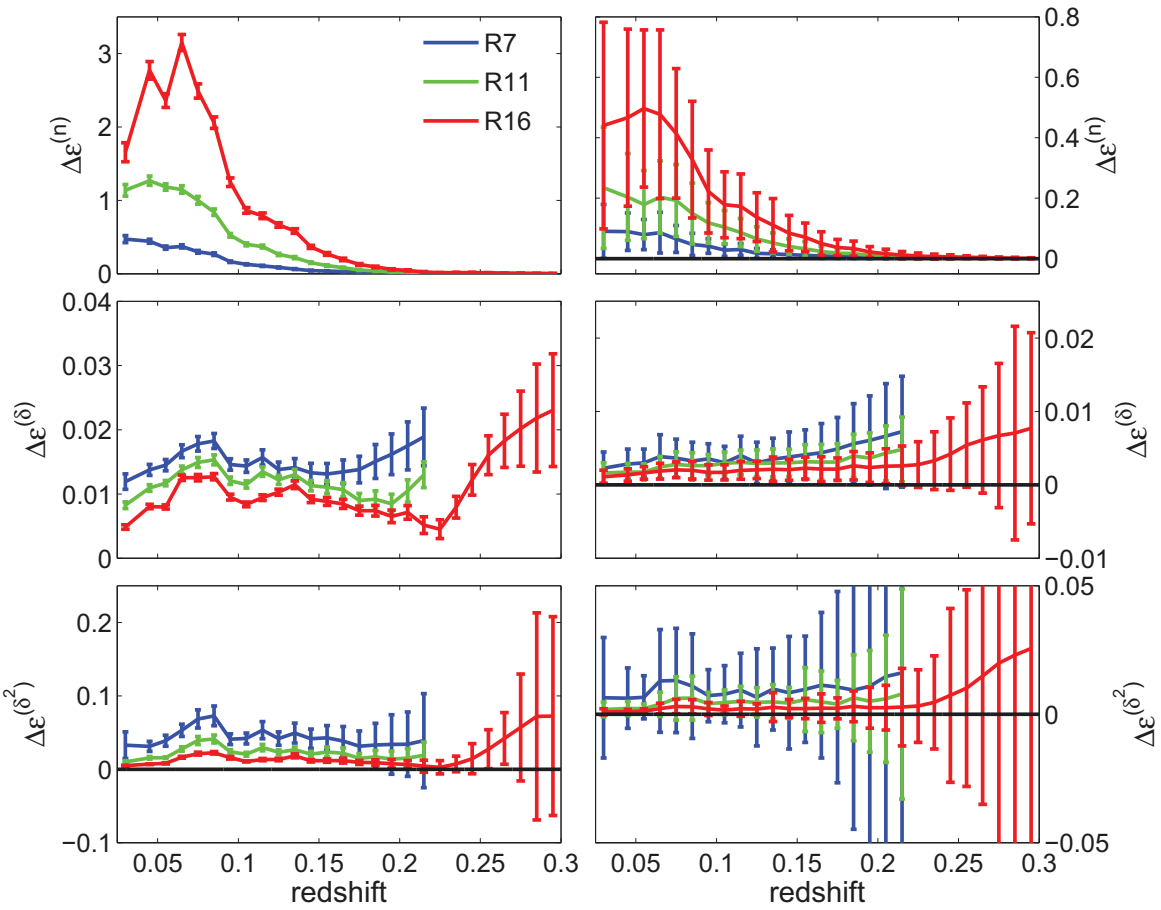

Figure 6.34: A comparison between selection function smearing and 2PCF smearing. For each cell size and redshift bin, $\epsilon^{(\delta)}$ for 2PCF smearing is subtracted from $\epsilon^{(\delta)}$ for selection function smearing to produce a comparison statistic $\Delta\epsilon^{(\delta)}$. When $\Delta\epsilon^{(\delta)} > 0$, 2PCF smearing is outperforming selection function smearing. Similar comparisons for $\Delta\epsilon^{(n)}$ and $\Delta\epsilon^{(\delta^2)}$ are made and presented in rows. The left and right columns contain the results for interspersed and dark regions respectively.

to each of the probabilistic methods. In this chapter, we have attempted to convey what we believe are the most relevant and interesting results, but we acknowledge that they are not comprehensive. For this reason, we have made the processed data files for each of the counting methods available online should the reader be compelled to explore the problem further.

6.4 Galaxies in Dark Regions

The dark region/object simulations of §6.3.3.1 made an important assumption — that dark regions are exclusively populated by *objects*. In practice, this is not always the case. The DR6 footprint corrections (see §5.2) occasionally recharacterize galaxies as lying *outside* the improved spectroscopic footprint. We refer to these MGS targets as *dark galaxies*.

The presence of *dark galaxies* has no impact on the counting of *dark objects* for all but one counting method — scaling. Naively applying the scaling relation of equation (6.2) necessarily discards information about the number of targets in the dark region. This simplification is worth avoiding if any of those targets' spectroscopic redshifts are known. However, simply adding the number count of *dark galaxies* to the approximated count of *dark objects* implies that the effective galaxy density in the dark region is possibly greater than in the rest of the cell.

Our solution is to derive an *effective spectroscopic completeness* β'_{spec} that better reflects the volume of the cell within the spectroscopic footprint once the *dark galaxies* are taken into account. The fundamental assumptions required to calculate β'_{spec} are 1) each target in the dark region occupies an equally-sized area of that dark region and 2) the area fraction occupied by each is the same as its volume fraction. Put another way, the surface density local to each target is assumed to be the same as that of the dark region as a whole.

To see this mathematically, consider a cell of volume V with a fraction β_{PS} inside the improved photometric footprint and a fraction β_{spec} within the improved spectroscopic footprint. Let n_{ig} equal the number of *interspersed galaxies* inside both the cell and the

spectroscopic footprint, n_{io} equal the number of *interspersed objects* approximated to be inside the cell (through the optimal interspersed counting strategy), n_{dg} equal the number of *dark galaxies* in the cell, n_{do} equal the number of *dark objects* within the cell's projection, and c equal the direction-dependent interspersed region spectroscopic completeness factor from equation (6.4).

The average fraction f of the cell's area (and volume, by assumption) occupied by *dark galaxies* and *dark objects* in the dark region is

$$f = \frac{\beta_{PS} - \beta_{spec}}{n_{dg} + n_{do}}. \tag{6.18}$$

These partial volumes are added to volume of the cell within the spectroscopic footprint such that

$$\beta'_{spec} = \beta_{spec} + f n_{dg}. \tag{6.19}$$

We replace β_{spec} in equation (6.3) with β'_{spec} from equation (6.19), and take the adjusted number of galaxies inside the cell's interspersed volume to be $n_{ig} + n_{io} + n_{dg}$. The number of targets n_t approximated to lie within the cell becomes

$$n_t = \left(\frac{\beta_{PS}(n_{dg} + n_{do})}{\beta_{spec} n_{do} + \beta_{PS} n_{dg}} \right) (n_{ig} + n_{io} + n_{dg}) \,. \tag{6.20}$$

Here is a summary of the scaling method when *dark galaxies* are present:

1. The number of interspersed galaxies n_{ig} are counted using equation (6.1).

2. The optimal interspersed counting method is applied to approximate the counts n_{io} of *interspersed objects*.

3. The average volume fraction f per dark region target is calculated and used to update the spectroscopic completeness volume in equation (6.19).

4. Using the number of *dark galaxies* n_{dg}, equation (6.20) is evaluated to deliver a final approximated number count.

Ultimately, the impact of *dark galaxies* is at the percent level. As illustrated in Figure 6.35, there are only limited number cells for which $\beta'_{spec} \neq \beta_{spec}$. For R7, R11, and R16 there are 366, 271, and 210 cells for which $\beta'_{spec} - \beta_{spec} \geq 0.002$, or 0.46%, 1.37%, and 1.38%, respectively.

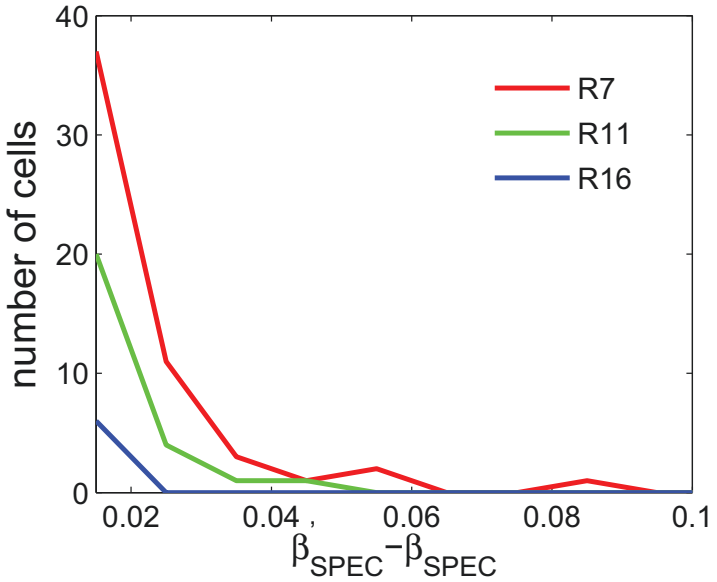

Figure 6.35: Distribution of cells for which $\beta'_{spec} - \beta_{spec} \geq 0.01$. Cells are counted in bins of width 0.01.

Consideration was given to the idea of treating *dark objects* whose nearest neighbors were within the spectroscopic footprint as *interspersed objects*. Their proximity to the footprint boundary might make them more akin to *interspersed objects* than to *dark objects*, which tend to clump together and lack galaxy neighbors on one side of their region.

However, none of the optimal interspersed methods involve the use of spatial correlations. Also, ignoring objects is preferred at high redshifts for both region types, further mitigating the need for a separate designation. While there could be some marginal benefit to scaling boundary objects rather than smearing them for low redshift R11 and R16 cells, the added complexity provides a disincentive. This particular issue is pursued no further.

6.5 Counting Results and Conclusions

This chapter investigated optimal counting strategies for n, δ, and δ^2 as a function of redshift, cell size, and region type. We found that the preferred techniques for approximating n and δ were almost identical, differing in only three instances and never significantly.

The optimal methods for δ^2 deviated somewhat from those of n and δ. These deviations were more likely when the optimal methods for n and δ changed rapidly with redshift, but were characteristically not of high significance. With interspersed regions, the deviations manifested themselves as a delay in preferring ignoring *objects* versus scaling them.

With dark regions, the probability of deviations increased with cell size. For R7 cells in the relatively limited redshift range of $0.06 < z < 0.09$, optimal counting of n and δ

demanded the scaling technique, while δ^2 called for probabilistic smearing. For R16 cells, there were disagreements between n/δ and δ^2 in 17 of the 19 redshift bins investigated between $0.04 < z < 0.23$. The R11 case fell in the middle — both in terms of the size of the redshift range of disagreement and the frequency of disagreements within that range.

The lack of total agreement among optimal n, δ, and δ^2 counting techniques begs the question of which counting method is truly best for a given cell size and redshift. For this, there can be no one absolute answer. Fortunately, the differences in the error metrics between optimal method candidates tend to be small, and agreement between measurement types is more likely than not. The disagreements that do exist were often the result of error metrics that differed to $< 1\sigma$ significance. This level of uncertainty suggests that there may be no singular answer that is "exactly right," and utilizing the preferred δ method over the preferred δ^2 method is unlikely to massively impact one's results.

Going forward, we utilize the optimal δ methods for a couple reasons. First, as we will demonstrate in Chapter 7, novel techniques exist to reduce systematic and shot noise from data when in the form of overdensities. It therefore makes sense to ensure raw measures of δ are as accurate as possible before attempting to correct them further. Moreover, assessing the impact of noise on large scale structure involves analyzing power spectra calculated using correlations of overdensities in cells.

Figure 6.36 displays the distribution of overdensities as a function of cell size. The set of R7 cells offers the highest resolution overdensity statistics by virtue of having the greatest number of cells. The large number of high-δ R7 cells results from many of those

cells having moderate $\langle n \rangle$, but large n. The R11 cells occupy the same redshift range as the R7 cells, but their larger size smooths out some of the structure due to a wider window function. The R16 cells are larger still, but over 56% lie in the range $0.22 < z \leq 0.30$ where $\langle n \rangle$ is very low and δ is high when galaxies are present.

In Figure 6.37, we report the percentage of the total MGS target count contributed by each galaxy and object type. For all cell sizes, interspersed galaxies comprise the largest percentage — one that at low z is only slightly smaller than the overall spectroscopic completeness of the DR6 survey. The interspersed fractions for galaxies and objects bend, respectively, towards one and zero at redshifts corresponding to the transition from scaling objects to ignoring them. A similar lurch towards zero for dark objects occurs at the redshift where ignoring objects becomes preferable.

The percentage of dark objects at low-z is $\sim 7\%$ for all cell sizes, a number that roughly reflects the average value of $1-\beta_{spec}$ in that range. Where dark objects are scaled (as in low-z R7 cells), they comprise a lower percentage of targets than when they are photometrically smeared (as in R16 cells). The percentage of dark objects in the R11 case, for which both scaling and smearing are used, lies in the middle.

We conclude this chapter by remarking that our application of these nine counting techniques has been very specific. They have been tested on the sixth Data Release of the Sloan Digital Sky Survey using only MGS targets selected within a predetermined photometric, spectroscopic, and angular range. Even so, the optimal counting methods still differed by cell size, region type, and measurement variety. These results will almost certainly differ

for other redshift surveys and cell geometries.

That said, this chapter should be viewed less as a universal statement on counting techniques and more as a general sketch. For instance, we can conclude that ignoring objects should be a favored option in regions where the selection function $\lesssim 10\%$ of its maximum value. We should also be open to using scaling, rather than nearest neighbor, to account for interspersed objects at low redshifts. The conclusion that 2PCF-smearing is almost always preferable to selection function smearing appears to be robust. It also seems that photometric redshift smearing would benefit from a shorter-tailed Gaussian or more individualized distribution. Beyond these broad strokes, individual conclusions can only be drawn by applying these testing methodologies to one's own data set.

We also emphasize that no individualized photo-z distributions per object are available on SkySever. Such distributions can, and have been, utilized to some success in other studies. Their existence suggests that neither probabilistic smearing nor even discrete photo-z's can be completely discounted. What has been reported here, therefore, should be thought of as the "floor" of what photometric redshifts can ultimately provide.

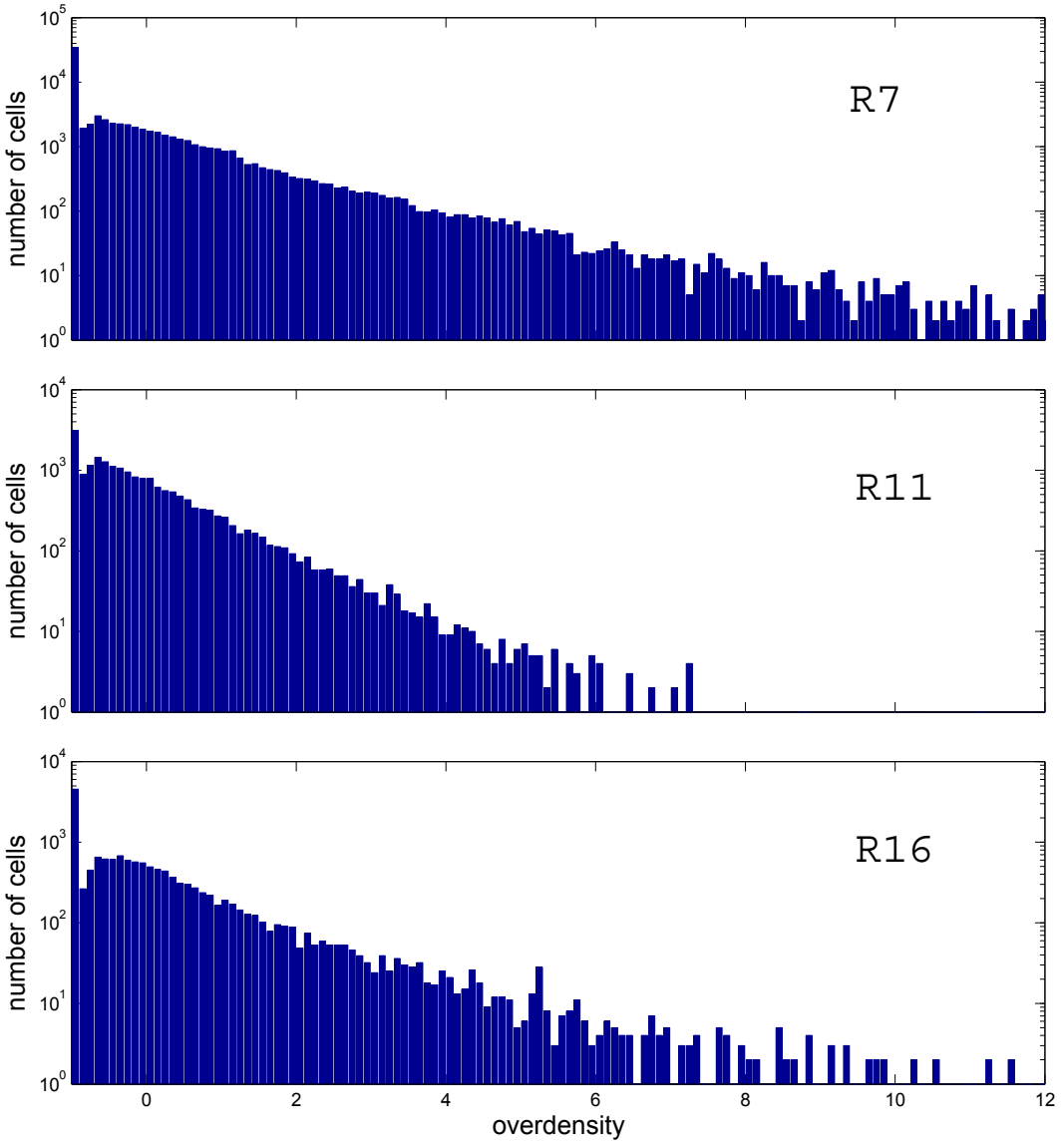

Figure 6.36: Histogram of MGS overdensities after the optimal δ counting techniques from Tables 6.1 and 6.2 are applied. Galaxies are counted in bins of width $\Delta\delta = 0.1$.

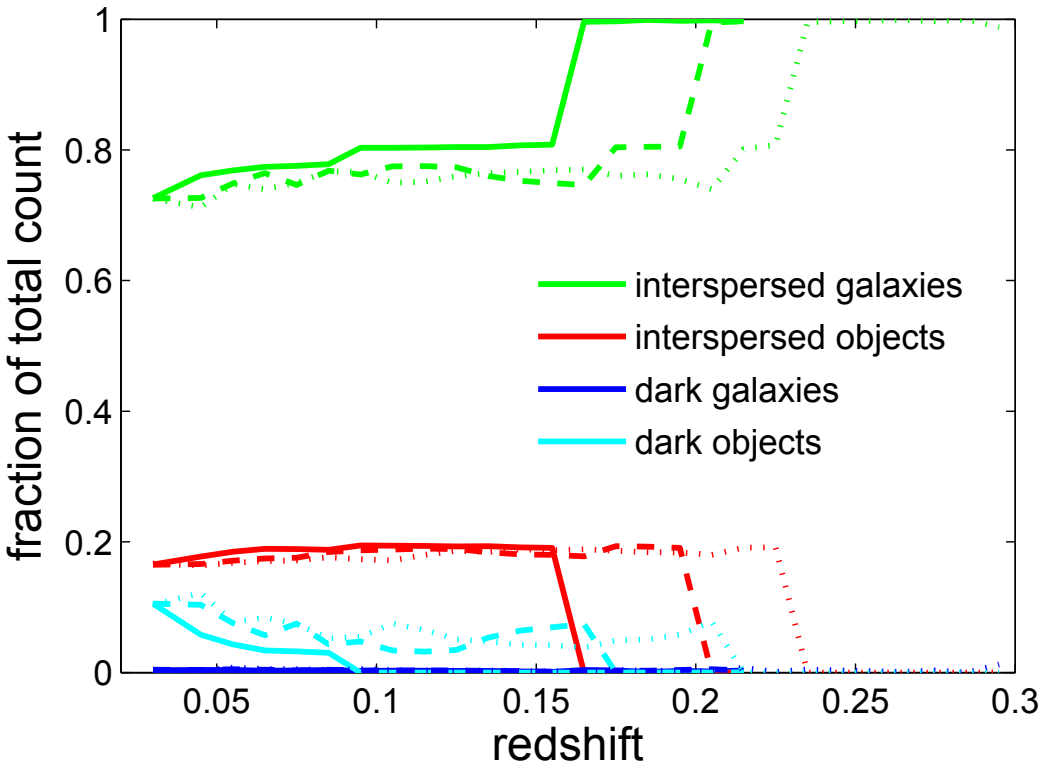

Figure 6.37: Fraction of total galaxy count n_t contributed by each of the four MGS target types. Counts are averaged over redshift bins of width $\Delta z = 0.01$. Results for R7 (*solid line*), R11 (*dashed line*), and R16 (*dotted line*) are presented.

Chapter 7

Data Cleansing – Theory

In this chapter, we introduce a method to predict and reduce the effect of shot noise and systematic errors in large data sets.

In general, a data vector can be represented as a sum of signal, shot noise, and systematic noise. Ideally, these would be disentangled from one another. In the pages that follow, we introduce analytic estimators for all three components in the case where both signal and noise are Gaussian. As discussed in Chapter 4, clustering overdensity, shot noise, and photometric zero-points all satisfy this condition.

We will show that, in measures of galactic overdensities, our framework can remove approximately 48% of shot and systematic noise variance introduced at low redshifts and over 82% at high redshifts. We also show that our estimated signal power spectrum is consistent with the true signal power spectrum and offers a significant improvement over taking the power of the raw data itself. Finally, we verify our analytic results empirically

using a Monte Carlo Markov Chain process and argue that its usefulness can be extended when signal and/or noise are non-Gaussian.

Before proceeding, the reader is encouraged to review §4.7 which describes the set of symbols used to represent signal, shot noise, and systematic noise in various spaces.

7.1 Expected Signal

The question our method seeks to answer is this: given the data $\boldsymbol{\delta}$, what is the most likely value of the underlying signal? An analytic solution exists when we assume the signal in cell-space is mean-zero and Gaussian with the distribution function,

$$P(\boldsymbol{\theta}) = \frac{\exp\left[-\frac{1}{2}\boldsymbol{\kappa}(\boldsymbol{\theta})^T \boldsymbol{\Sigma}_\kappa^{-1} \boldsymbol{\kappa}(\boldsymbol{\theta})\right]}{\sqrt{(2\pi)^N \left|\boldsymbol{\Sigma}_\kappa\right|}}. \tag{7.1}$$

If the noise is Gaussian, $\boldsymbol{\delta}$ is symmetric about the signal. The probability of obtaining a particular data vector given a set of signal parameters $\boldsymbol{\theta}$ would then be

$$P(\boldsymbol{\delta}|\boldsymbol{\theta}) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\kappa}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_{\eta\zeta}^{-1} (\boldsymbol{\delta} - \boldsymbol{\kappa}(\boldsymbol{\theta}))\right]}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}_{\eta\zeta}|}}. \tag{7.2}$$

The expected value of the i^{th} signal coefficient can be expressed as a function of the posterior probability of signal parameters $\boldsymbol{\theta}$ given the data,

$$\langle \kappa_i | \boldsymbol{\delta} \rangle = \int \kappa_i(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\delta})\, d\boldsymbol{\theta}. \tag{7.3}$$

The posterior probability is most effectively solved for using Bayes's theorem, $P(\boldsymbol{\theta}|\boldsymbol{\delta}) = P(\boldsymbol{\delta}|\boldsymbol{\theta})P(\boldsymbol{\theta})/P(\boldsymbol{\delta})$. To simplify the notation we define

$$c \equiv \frac{\exp\left[-\frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\eta\zeta}^{-1}\boldsymbol{\delta}\right]}{(2\pi)^N\sqrt{|\boldsymbol{\Sigma}_\kappa||\boldsymbol{\Sigma}_{\eta\zeta}|}}. \tag{7.4}$$

The numerator of Bayes's theorem equals

$$P(\boldsymbol{\delta}|\boldsymbol{\theta})P(\boldsymbol{\theta}) = c\exp\left[-\frac{1}{2}(-2\boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\eta\zeta}^{-1}\boldsymbol{\kappa}(\boldsymbol{\theta}) + \boldsymbol{\kappa}(\boldsymbol{\theta})^T(\boldsymbol{\Sigma}_\kappa^{-1} - \boldsymbol{\Sigma}_{\eta\zeta}^{-1})\boldsymbol{\kappa}(\boldsymbol{\theta})\right]. \tag{7.5}$$

The sum of inverse covariance matrices can be diagonalized via

$$\boldsymbol{\Sigma}_\kappa^{-1} + \boldsymbol{\Sigma}_{\eta\zeta}^{-1} = \mathbf{W}\boldsymbol{\Lambda}^{(W)}\mathbf{W}^T. \tag{7.6}$$

Because $\mathbf{W}$ is orthonormal, $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, and equation (7.5) can be recast as

$$P(\boldsymbol{\delta}|\boldsymbol{\theta})P(\boldsymbol{\theta}) = c\exp\left[\boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\eta\zeta}^{-1}\mathbf{W}\mathbf{W}^T\boldsymbol{\kappa}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\kappa}(\boldsymbol{\theta})^T\mathbf{W}\mathbf{W}^T\left(\boldsymbol{\Sigma}_\kappa^{-1} + \boldsymbol{\Sigma}_{\eta\zeta}^{-1}\right)\mathbf{W}\mathbf{W}^T\boldsymbol{\kappa}(\boldsymbol{\theta})\right]. \tag{7.7}$$

By virtue of combining $\boldsymbol{s} = \mathbf{W}^T\boldsymbol{\kappa}$ with equation (7.7),

$$P(\boldsymbol{\delta}|\boldsymbol{\theta})P(\boldsymbol{\theta}) = c\exp\left[\boldsymbol{x}^T\boldsymbol{s}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{s}(\boldsymbol{\theta})^T\boldsymbol{\Lambda}^{(W)}\boldsymbol{s}(\boldsymbol{\theta})\right], \tag{7.8}$$

where $\boldsymbol{x}^T \equiv \boldsymbol{\delta}^T \boldsymbol{\Sigma}_{\eta\zeta}^{-1} \mathbf{W}$. To normalize this function we note that $\int P(\boldsymbol{\theta}|\boldsymbol{\delta})\, d\boldsymbol{\theta} = 1$ and therefore $P(\boldsymbol{\delta}) = \int P(\boldsymbol{\delta}|\boldsymbol{\theta}) P(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

We have the freedom to choose any N parameters that most conveniently map $\boldsymbol{\theta}$ to the signal. The form of the probability distribution function suggests a natural choice of $\theta_i = s_i \; \forall\, i$. From this perspective, it is more convenient to solve the problem

$$\langle s_i|\boldsymbol{\delta}\rangle = \int s_i\, P(\boldsymbol{s}|\boldsymbol{\delta})\, d\boldsymbol{s}. \tag{7.9}$$

Because of the diagonality of $\boldsymbol{\Lambda}^{(W)}$, the expression for $P(\boldsymbol{\delta}|\boldsymbol{\theta})P(\boldsymbol{\theta})$ is separable in the exponent,

$$\begin{aligned}
\langle s_i|\boldsymbol{\delta}\rangle &= \frac{\int_{-\infty}^{\infty} s_i\, P(\boldsymbol{\delta}|\boldsymbol{s}) P(\boldsymbol{s})\, d\boldsymbol{s}}{\int_{-\infty}^{\infty} P(\boldsymbol{\delta}|\boldsymbol{s}') P(\boldsymbol{s}')\, d\boldsymbol{s}'} \\
&= \frac{\int_{-\infty}^{\infty} s_i \exp\left[\boldsymbol{x}^T \boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T \boldsymbol{\Lambda}^{(W)} \boldsymbol{s}\right] d\boldsymbol{s}}{\int_{-\infty}^{\infty} \exp\left[\boldsymbol{x}^T \boldsymbol{s}' - \frac{1}{2}\boldsymbol{s}'^T \boldsymbol{\Lambda}^{(W)} \boldsymbol{s}'\right] d\boldsymbol{s}'} \\
&= \frac{\int_{-\infty}^{\infty} s_i \prod_{j=1}^{N} \exp\left[x_j s_j - \frac{1}{2} s_j^2 \lambda_j^{(W)}\right] ds_j}{\int_{-\infty}^{\infty} \prod_{j=1}^{N} \exp\left[x_j s_j' - \frac{1}{2} s_j'^2 \lambda_j^{(W)}\right] ds_j'}.
\end{aligned} \tag{7.10}$$

All but one of the integrals in the numerator and denominator of equation (7.10) cancels.

$$\langle s_i|\boldsymbol{\delta}\rangle = \frac{\int_{-\infty}^{\infty} s_i \exp\left[x_i s_i - \frac{1}{2}s_i^2\lambda_i^{(W)}\right] ds_i}{\int_{-\infty}^{\infty} \exp\left[x_i s_i' - \frac{1}{2}s_i'^2\lambda_i^{(W)}\right] ds_i'}$$
$$= x_i/\lambda_i^{(W)}. \tag{7.11}$$

Once all values of $\boldsymbol{s}$ are evaluated, the solution in cell-space can be found through one final rotation, $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle = \mathbf{W}\langle\boldsymbol{s}|\boldsymbol{\delta}\rangle$.

7.2 Expected Shot Noise

The derivation for the expected shot noise given the data is similar to that of the expected signal. We assume the shot noise $\zeta(\varphi)$ is mean-zero and Gaussian with the distribution function,

$$P(\boldsymbol{\varphi}) \propto \exp\left[-\frac{1}{2}\boldsymbol{\zeta}(\boldsymbol{\varphi})^T\boldsymbol{\Sigma}_\zeta^{-1}\boldsymbol{\zeta}(\boldsymbol{\varphi})\right]. \tag{7.12}$$

The signal plus systematic noise is also Gaussian, therefore $\boldsymbol{\delta}$ is symmetric around $\boldsymbol{\zeta}$ such that the probability of obtaining a particular data vector given a set of shot noise parameters $\boldsymbol{\varphi}$ is

$$P(\boldsymbol{\delta}|\boldsymbol{\varphi})P(\boldsymbol{\varphi}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\delta}-\boldsymbol{\zeta}(\boldsymbol{\varphi}))^T\boldsymbol{\Sigma}_{\kappa\eta}^{-1}(\boldsymbol{\delta}-\boldsymbol{\zeta}(\boldsymbol{\varphi})\right]. \tag{7.13}$$

Taking the product of the probabilities,

$$\begin{aligned} P(\boldsymbol{\delta}|\boldsymbol{\varphi})P(\boldsymbol{\varphi}) &\propto \exp\left[\boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\kappa\eta}^{-1}\boldsymbol{\zeta}(\boldsymbol{\varphi})\right]\exp\left[-\frac{1}{2}\boldsymbol{\zeta}(\boldsymbol{\varphi})^T(\boldsymbol{\Sigma}_{\zeta}^{-1}+\boldsymbol{\Sigma}_{\kappa\eta}^{-1})\boldsymbol{\zeta}(\boldsymbol{\varphi})\right] \\ &\propto \exp\left[\boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\kappa\eta}^{-1}\mathbf{B}\boldsymbol{\pi}(\boldsymbol{\varphi})\right]\exp\left[-\frac{1}{2}\boldsymbol{\pi}(\boldsymbol{\varphi})^T\boldsymbol{\Lambda}^{(B)}\boldsymbol{\pi}(\boldsymbol{\varphi})\right] \\ &\propto \exp\left[\boldsymbol{h}^T\boldsymbol{\pi}(\boldsymbol{\varphi})-\frac{1}{2}\boldsymbol{\pi}(\boldsymbol{\varphi})^T\boldsymbol{\Lambda}^{(B)}\boldsymbol{\pi}(\boldsymbol{\varphi})\right], \end{aligned} \tag{7.14}$$

where $\boldsymbol{h}^T = \boldsymbol{\delta}^T\boldsymbol{\Sigma}_{\kappa\eta}^{-1}\mathbf{B}$ and the sum of inverse covariance matrices is diagonalized via

$$\boldsymbol{\Sigma}_{\zeta}^{-1}+\boldsymbol{\Sigma}_{\kappa\eta}^{-1} = \mathbf{B}\boldsymbol{\Lambda}^{(B)}\mathbf{B}^T. \tag{7.15}$$

Changing the magnitude σ_m of the zero-point noise coefficients requires a complete reevaluation of $\boldsymbol{\Sigma}_{\eta\zeta}^{-1}$. Being a full-rank matrix, this inversion is very time consuming. However, the limited dimensionality of the zero-point noise enables the use of a mathematical shortcut, which we describe in Appendix H.

Choosing $\boldsymbol{\varphi} = \boldsymbol{\pi}$ as the shot noise parameters, the expected value of the i^{th} shot noise coefficient in B-space may calculated using equation (7.16). The second equality comes from Bayes's theorem.

$$\langle\pi_i|\boldsymbol{\delta}\rangle = \int \pi_i P(\boldsymbol{\pi}|\boldsymbol{\delta})d\boldsymbol{\pi} = \int \pi_i P(\boldsymbol{\delta}|\boldsymbol{\pi})P(\boldsymbol{\pi})d\boldsymbol{\pi}. \tag{7.16}$$

Following a similar argument to that employed for W-space, we find

$$\langle \pi_i | \boldsymbol{\delta} \rangle = \frac{h_i}{\lambda_i^{(B)}}. \tag{7.17}$$

Finally, rotate the result back into cell-space with $\langle \boldsymbol{\zeta} | \boldsymbol{\delta} \rangle = \mathbf{B} \langle \boldsymbol{\pi} | \boldsymbol{\delta} \rangle$.

7.3 Expected Zero-Point Noise

While it is tempting to repeat the derivations of Sections 7.1 and 7.2 for the systematic noise, it is non-optimal if those previous two solutions are already known. This is the case for two reasons. First, taking a diagonalization of the kind represented in equation (7.6) is expensive and ought to be avoided if possible. Second, while signal and shot noise are often full rank, especially in cosmology, systematic noise in general has fewer degrees of freedom m than there are cells N.

When $\boldsymbol{\Sigma}_\eta = \mathbf{U} \boldsymbol{\Lambda}^{(\eta)} \mathbf{U}^T$ has a rank $m < N$, $N - m$ diagonal elements of $\boldsymbol{\Lambda}^{(\eta)}$ equal zero. This makes it impossible to directly evaulate $\boldsymbol{\Sigma}_\eta^{-1} = \mathbf{U} \boldsymbol{\Lambda}^{(\eta)^{-1}} \mathbf{U}^T$ since $N - m$ of the diagonal elements of $\boldsymbol{\Lambda}^{(\eta)^{-1}}$ equal infinity. The simpler approach is to observe that since $\boldsymbol{\delta} = \boldsymbol{\kappa} + \boldsymbol{\eta} + \boldsymbol{\zeta}$,

$$\langle \boldsymbol{\delta} | \boldsymbol{\delta} \rangle = \langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle + \langle \boldsymbol{\eta} | \boldsymbol{\delta} \rangle + \langle \boldsymbol{\zeta} | \boldsymbol{\delta} \rangle, \tag{7.18}$$

and thus,

$$\langle \boldsymbol{\eta}|\boldsymbol{\delta}\rangle = \boldsymbol{\delta} - \langle \boldsymbol{\kappa}|\boldsymbol{\delta}\rangle - \langle \boldsymbol{\zeta}|\boldsymbol{\delta}\rangle. \tag{7.19}$$

7.4 Empirical Verification

The results of §7.1 can be verified empirically using a Monte Carlo process. If $\boldsymbol{\theta}^{(\tau)}|\boldsymbol{\delta}$ represents random variates drawn from the distribution $P(\boldsymbol{\theta}|\boldsymbol{\delta})$, then the estimated signal given the data in cell-space is

$$\hat{\kappa}_i = \frac{1}{K}\sum_{\tau=1}^{K} \kappa_i(\boldsymbol{\theta}^{(\tau)}|\boldsymbol{\delta}), \tag{7.20}$$

where $\langle \kappa_i|\boldsymbol{\delta}\rangle = \lim_{K\to\infty} \hat{\kappa}_i$. To generate random variates from the distribution $P(\boldsymbol{\theta}|\boldsymbol{\delta})$ we employ a Monte Carlo Markov Chain (MCMC) method. Using the Metropolis-Hastings algorithm (for a full treatment see Bolstad, 2012, p. 130), we can draw variates provided we possess a function that is proportional to that distribution. As shown in equation (7.8), we have one such function already at our disposal,

$$g(\boldsymbol{\theta}|\boldsymbol{\delta}) = \exp\left[\boldsymbol{x}^T \boldsymbol{s}(\boldsymbol{\theta})\right] \exp\left[-\frac{1}{2}\boldsymbol{s}(\boldsymbol{\theta})^T \boldsymbol{\Lambda}^{(W)} \boldsymbol{s}(\boldsymbol{\theta})\right]. \tag{7.21}$$

We must also introduce an *independent candidate density* $q(\boldsymbol{\theta})$ that will ideally match the correlation structure of $g(\boldsymbol{\theta}|\boldsymbol{\delta})$ but with a broader tail so that the former "blankets" the latter. But first, the steps for the Metropolis-Hastings method are as follows:

1. Choose an initial set of signal parameters $\boldsymbol{\theta}^{(0)}$ that preferably lies near the peak of $g(\boldsymbol{\theta}|\boldsymbol{\delta})$.

2. Repeat the following steps n times for $i = 1, \ldots, n$:

   (a) Draw a new random vector $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta})$.

   (b) Calculate the *acceptance probability* $\beta(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}')$ where

$$\beta(\boldsymbol{\theta}|\boldsymbol{\theta}') = \min\left[1, \frac{g(\boldsymbol{\theta}'|\boldsymbol{\delta})q(\boldsymbol{\theta})}{g(\boldsymbol{\theta}|\boldsymbol{\delta})q(\boldsymbol{\theta}')}\right], \tag{7.22}$$

   (c) Draw a random variate u from the uniform distribution $U(0, 1)$.

   (d) If $u < \beta(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}')$, then set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'$ and accept $\boldsymbol{\theta}'$ as one of the random variates of $P(\boldsymbol{\theta}|\boldsymbol{\delta})$. Otherwise, set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$.

Our goal in choosing an independent candidate density $q(\boldsymbol{\theta})$ is finding one with approximately the same shape as the target distribution $g(\boldsymbol{\theta}|\boldsymbol{\delta})$ and from which we can draw random vectors. A reasonable first guess is a multivariate Gaussian distribution that peaks in the same place as $g(\boldsymbol{\theta}|\boldsymbol{\delta})$—let's call it $\hat{\boldsymbol{\theta}}$—and has the same curvature as $g(\boldsymbol{\theta}|\boldsymbol{\delta})$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. However, in practice the Gaussian is usually not broad enough to provide adequate coverage at the tails of $g(\boldsymbol{\theta}|\boldsymbol{\delta})$.

A better candidate density is the multivariate t-distribution $t_f(\hat{\boldsymbol{\theta}}, \mathbf{Y})$ which allows for more representative sampling at low degrees of freedom f when $\mathbf{H}$ is the curvature at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. (As f increases, t_f thins and asymptotically approaches the multivariate Gaussian

distribution—it is best to not go too far in this direction.)

To find the peak and curvature we begin with the logarithm of the target density,

$$l(\boldsymbol{\theta}|\boldsymbol{\delta}) = \ln(g(\boldsymbol{\theta}|\boldsymbol{\delta})) = \boldsymbol{x}^T \boldsymbol{s}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{s}(\boldsymbol{\theta})^T \boldsymbol{\Lambda}^{(W)} \boldsymbol{s}(\boldsymbol{\theta}). \tag{7.23}$$

We note that a function and its logarithm have maxima located at the same position. The first derivative of $l(\boldsymbol{\theta}|\boldsymbol{\delta})$ where $\theta_i = s_i$ is

$$\begin{pmatrix} \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_N} \end{pmatrix} = \begin{pmatrix} x_1 - \lambda_1^{(W)} s_1 \\ \vdots \\ x_N - \lambda_N^{(W)} s_N \end{pmatrix}. \tag{7.24}$$

Setting equation (7.24) equal to zero reveals the location of the target density's maximum. This happens to be the explicit analytic solution of the integral we found in equation (7.11).

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} x_1/\lambda_1^{(W)} \\ \vdots \\ x_N/\lambda_N^{(W)} \end{pmatrix}. \tag{7.25}$$

The inverse of the target density's second derivative at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ yields the curvature of a multivariate normal that peaks at the same location,

$$\begin{aligned}\mathbf{H} &= -\left[\begin{array}{ccc} \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_1^2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_1 \partial \theta_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_N \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\delta})}{\partial \theta_N^2} \end{array}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= -\left[\begin{array}{ccc} -\lambda_1^{(W)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\lambda_N^{(W)} \end{array}\right]^{-1}. \end{aligned} \tag{7.26}$$

Equation (7.26) simplifies to $\mathbf{H} = \boldsymbol{\Lambda}^{(W)^{-1}}$. Dropping the constant terms (since they cancel in $\beta\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right)$), the independent candidate distribution adopts the form of $t_f\left(\hat{\boldsymbol{\theta}}, \mathbf{H}\right)$,

$$q\left(\boldsymbol{\theta}\right) = \left(1 + \frac{1}{f}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \boldsymbol{\Lambda}^{(W)} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right)^{-(N+f)/2}. \tag{7.27}$$

To sample $\boldsymbol{\theta}'$ from this distribution, solve for the lower triangular matrix $\mathbf{L}$ that satisfies the Cholesky decomposition $\mathbf{H} = \mathbf{L}\mathbf{L}^T$. The diagonality of $\mathbf{H}$ yields a simple solution,

$$\mathbf{L} = \left[\begin{array}{ccc} 1/\sqrt{\lambda_1^{(W)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sqrt{\lambda_N^{(W)}} \end{array}\right]. \tag{7.28}$$

When a set of random variables $\tilde{t} = \left(\tilde{\boldsymbol{t}}_1 \cdots \tilde{t}_N\right)^T$ is drawn from the t-distribution with $f = 1$, the vector $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \mathbf{L}\tilde{\boldsymbol{t}}$ will constitute a random draw from $q\left(\boldsymbol{\theta}\right)$.

At this point the benefits of working in W-space should be clear. Every Metropolis-Hastings draw requires evaluation of equations (7.21) and (7.27). By casting them in terms of $\mathbf{\Lambda}^{(W)}$, an expensive pair of matrix-vector products is transformed into a relatively simple summation. When the number of MCMC draws needed for convergence is large, this computational simplification can lead to enormous performance gains.

The Metropolis-Hastings algorithm works by always accepting a variate that is "uphill" of its current position and selecting one that is "downhill" with a probability β. Accepting too many variates oversamples the low probability regions of the t-distribution while accepting too few can needlessly lengthen the time needed for representative sampling.

A proper balance is struck by calibrating the number of degrees of freedom f. Roberts et al. (1997) found the ideal acceptance rate for an N-dimensional Gaussian to be about 23% when N samples are drawn. By varying the degrees of freedom and observing the acceptance rate, it is straightforward to determine the optimal f for the Metropolis-Hastings draws.

Draws should continue until convergence is achieved. To test this, rotate each $\boldsymbol{s}^{(\tau)}$ back into cell-space through $\boldsymbol{\kappa}^{(\tau)} = \mathbf{W}\boldsymbol{s}^{(\tau)}$ then plot the signal estimates as a function of the number of realizations R using $\hat{\kappa}_i = (1/R)\sum_{\tau=1}^{R} \kappa_i^{(\tau)}$.

However, storing all $\boldsymbol{s}^{(\tau)}$, rotating them through $\mathbf{W}$, and processing all $\boldsymbol{\kappa}^{(\tau)}$ can be an expensive operation in terms of disk space, memory, and CPU. If the expected signal vector $\hat{\boldsymbol{\kappa}}$ is the only deliverable needed, then a mathematical shortcut can be employed by only storing the sum of $\boldsymbol{s}^{(\tau)}|\boldsymbol{\delta}$,

$$\hat{\boldsymbol{\kappa}} = \frac{1}{K}\sum_{\tau=1}^{K} \mathbf{W}\left(\boldsymbol{s}^{(\tau)}|\boldsymbol{\delta}\right) = \mathbf{W}\left(\frac{1}{K}\sum_{\tau=1}^{K}\left(\boldsymbol{s}^{(\tau)}|\boldsymbol{\delta}\right)\right). \tag{7.29}$$

Finally, while this derivation assumed Gaussian signal and noise, other distributions may be substituted. The particular mathematics will likely change, and to maximize speed a new version of W-space might need to be constructed, but in principle the problem should still be soluble. Being an empirical process, the Metropolis-Hastings approach is necessarily slower than the analytic one by several orders of magnitude. But in the event the latter does not exist, this MCMC method should be considered as a viable alternative.

7.5 Signal Estimation

In this section we apply the signal estimation method of §7.1 to the problem of reducing statistical and systematic noise from MGS overdensity data. Using the steps outlined in Chapter 4, we generate random instances of clustering signal, zero-point noise and shot noise overdensity vectors $\boldsymbol{\delta}_\kappa$, $\boldsymbol{\delta}_\eta$, and $\boldsymbol{\delta}_\zeta$. For the purpose of our diagnostic tests these vectors remain "hidden" and only their sum $\boldsymbol{\delta}$ is known to our cleansing algorithm.

7.5.1 Cell Statistics

To begin, individual $\boldsymbol{\delta}$ vectors are processed through equation (7.11) to yield signal estimates $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$. In Figure 7.1 we examine each cell individually to assess the effect of

the cleansing on a single overdensity realization. The values plotted on the vertical and horizontal axes plot the distance between the true signal $\boldsymbol{\delta}_\kappa$ and, respectively, the raw data $\boldsymbol{\delta}$ and the reconstructed signal $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$. If the cleansing was perfect, all points would be aligned vertically at 0, indicating that the difference between the signal estimate and the signal had been reduced to zero. It should be considered a success if points lie above the blue line, i.e. if the cleansing process yields a signal estimate that is closer to the actual signal than the data itself is.

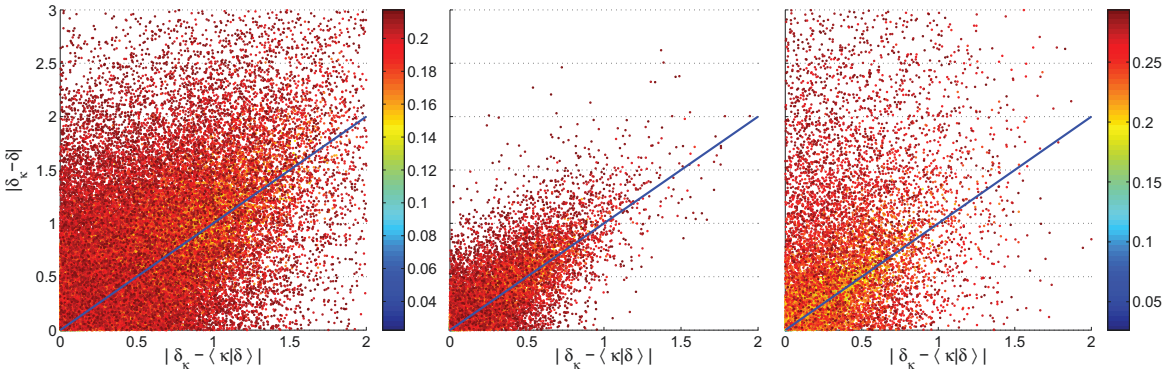

Figure 7.1: Response of a simulated data vector $\boldsymbol{\delta}$ to signal estimation. Each pixel represents a single cell in R7 (*left*), R11 (*middle*), and R16 (*right*) with color-coded redshift. The vertical axis represents the distance between an element's signal component $\delta_{\kappa,i}$ and its data component δ_i, the latter of which also has zero-point noise and shot noise added in. The horizontal axis contains that same measure but between $\delta_{\kappa,i}$ and the reconstructed signal vector $\langle \kappa_i | \boldsymbol{\delta} \rangle$. The left color bar is for the R7 and R11 cases, while the right is for R16. The blue line has unit slope.

The lower redshift objects, marked in blues and greens, are huddled around the blue line, indicating that they were largely unaffected by the cleansing process. This is entirely expected since the zero-point noise and shot noise only begin to play a significant role on galaxy counts at larger redshifts. We observe that the greatest amounts of scatter, and the greatest improvements, occur within the highest redshift cells.

Radius (h^{-1}Mpc)	Cells Impacted Positively	$\|\|\boldsymbol{\kappa}^{(\tau)} - \boldsymbol{\Gamma}^{(\tau)}\|\|_2$	$\|\|\boldsymbol{\kappa}^{(\tau)} - \langle\boldsymbol{\kappa}\|\boldsymbol{\Gamma}^{(\tau)}\rangle\|\|_2$	Drop in noise variance
7	$60.4 \pm 0.2\%$	$234.3 \pm 0.8\%$	$169.1 \pm 0.5\%$	$47.9 \pm 0.4\%$
11	$57.1 \pm 0.4\%$	$58.0 \pm 0.4\%$	$48.7 \pm 0.3\%$	$29.5 \pm 0.8\%$
16	$65.1 \pm 0.4\%$	$127.5 \pm 1.3\%$	$53.7 \pm 0.4\%$	$82.3 \pm 0.4\%$

Table 7.1: Average response of 10,000 random overdensity data vectors to cleansing. Cells are defined as being positively impacted if the distance between the signal and reconstructed signal is smaller than that between the signal and the data.

We quantify this process in aggregate by estimating the signal $\langle\boldsymbol{\kappa}|\boldsymbol{\Gamma}^{(\tau)}\rangle$ for 10,000 simulated data vectors $\boldsymbol{\Gamma}^{(\tau)}$. We examined each cell's updated signal value $\langle\kappa_i|\boldsymbol{\Gamma}^{(\tau)}\rangle$ and found that for all sphere sizes a majority of cells benefited from the cleansing, i.e. $|\kappa_i^{(\tau)} - \langle\kappa_i|\boldsymbol{\Gamma}^{(\tau)}\rangle| < |\kappa_i^{(\tau)} - \Gamma_i^{(\tau)}|$. We also measure the total distance of $\langle\boldsymbol{\kappa}|\boldsymbol{\Gamma}^{(\tau)}\rangle$ and $\boldsymbol{\Gamma}^{(\tau)}$ from the true signal $\boldsymbol{\kappa}_{(\tau)}$ using the vector 2-norm. These results are summarized in Table **??**.

Noise variance drops more for R7 than R11 for two reasons. First, the R7 spheres, by having smaller circular projections on the sky, contain fewer SEGMENTs. This reduces the potentially offsetting effects of adjacent zero-points and allows for more accurate probing of individual offsets.

Second, because R7 cells have smaller volumes, they also have lower expected numbers of galaxies and consequently higher shot noise. The cleansing algorithm has identified that shot noise, leading to a larger correction. A similar effect is in play with the R16 cells, though for a different reason. Here $\langle n\rangle$ values are smaller because the cells extend to high redshifts as opposed to having smaller sizes. All else being equal, within a fixed volume a greater percentage of noise is reduced if more cells are used.

Figure 7.2 illustrates that most of the distance between $\boldsymbol{\kappa}^{(\tau)}$ and $\boldsymbol{\Gamma}^{(\tau)}$ accumulates at

large redshifts. This is due to both the greater number of cells present there as well as the fact that shot and zero-point noise play a greater role at high z. When $\boldsymbol{\Gamma}^{(\tau)}$ is replaced with $\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ the distance still increases with redshift, but more slowly. So while the noise has its greatest impact at large redshifts, so too does the cleansing procedure.

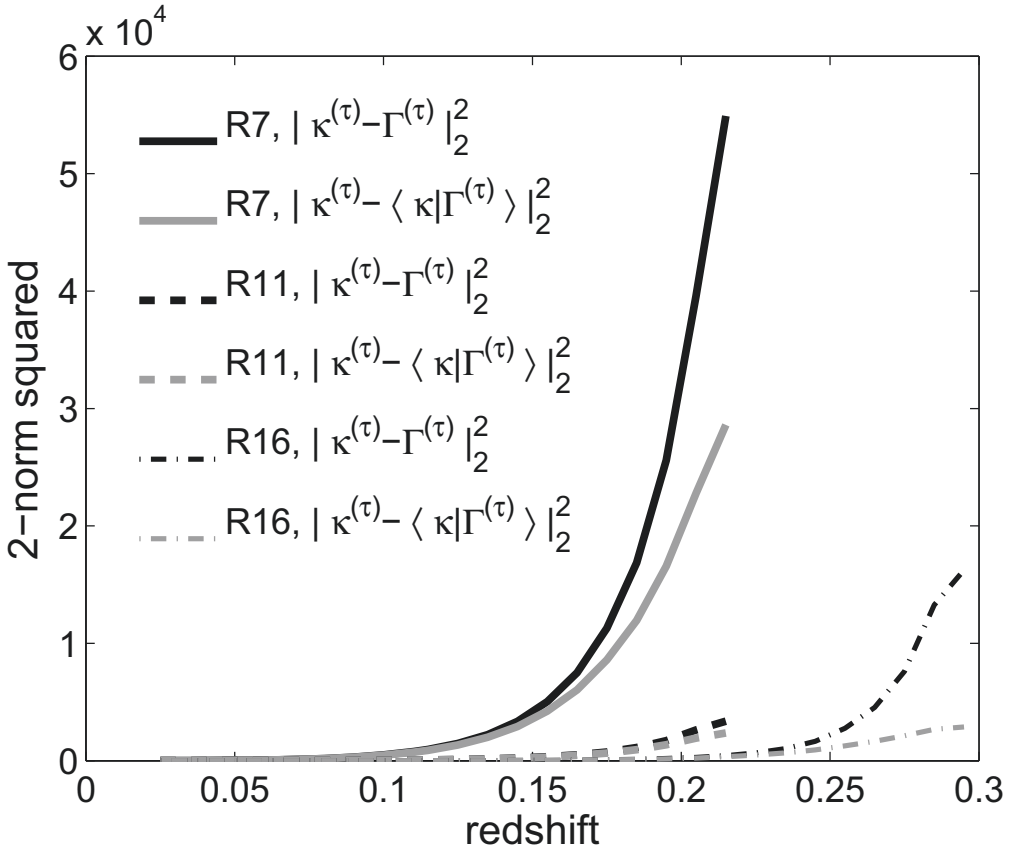

Figure 7.2: The cumulative variance of the distance from the signal is shown as a function of redshift when $\sigma_m = 0.01$. Results shown are the averages over 2000 realizations. The plot is divided into 20 equally-spaced redshift bins for R7 and R11 and 28 bins for R16. The top line of each radius pair follows the two-norm distance squared between $\boldsymbol{\kappa}^{(\tau)}$ and $\boldsymbol{\Gamma}^{(\tau)}$ using only cells at or below the given redshift. The bottom line of each pair traces the cumulative variance of $\boldsymbol{\kappa}^{(\tau)} - \langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$.

7.5.2 Power Spectra

We examine the power spectra of the signal $\boldsymbol{\kappa}$, shot noise $\boldsymbol{\zeta}$, zero-point noise $\boldsymbol{\eta}$, data $\boldsymbol{\Gamma}$ (which is the sum of the previous 3 components) and estimated signal $\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ using the non-parametric estimator introduced in §3.2.

The power spectra of the data components are presented in Figure 7.3. The power of the raw data is necessarily larger than each of its constituent parts at all wavenumbers. The largest difference between the signal and data is shot noise. The shot noise is largely flat as this white noise contributes roughly the same power on all scales. Its amplitude is slightly larger for R7 than R11, but R7's is significantly smaller than R16's whose cells extend to larger redshifts where the number of expected galaxies is lower. Signal power is uniform across cell sizes except at large k where the smearing effect of the cells causes a drop in power for it and the noise power.

The zero-point noise power is highest for R16 since $f(z)$ is greatest at high redshifts. This power is partially mitigated by the size of its cells. Larger cells enclose more PRIMARY SEGMENTs on average, permitting zero-points to offset, though this effect plays a diminishing role at higher z. Figure 7.3 reveals that of these two the rate of change in the selection function is the dominant process.

We can estimate the zero-point spectrum a priori provided we understand something about the correlation function of the noise. As covered in §6.2, we often approximate the correlation function $\xi(r)$ as the ratio of the number of objects observed to the number expected, at some separation distance r. In a three-dimensional space centered at any point,

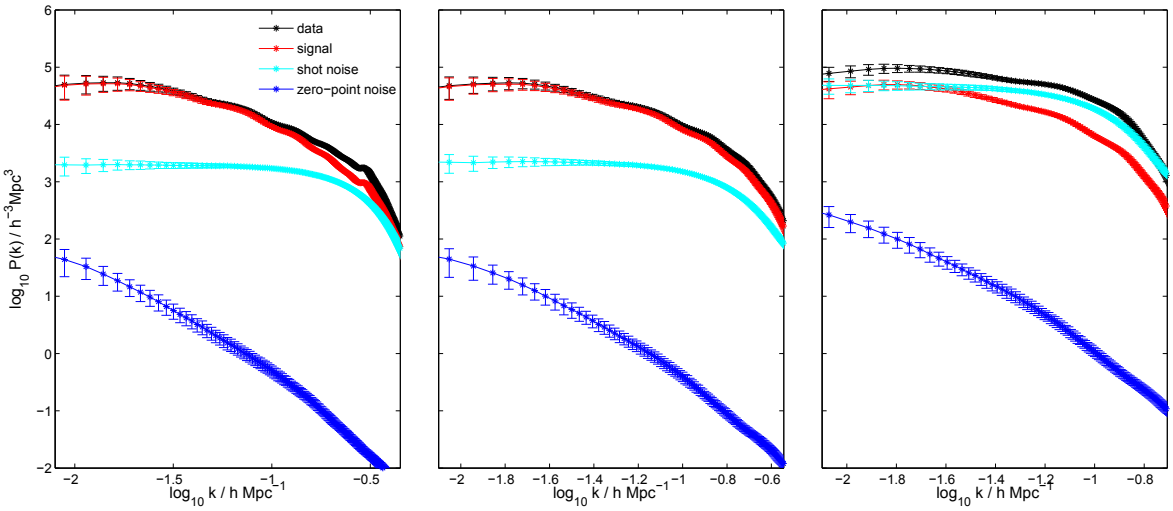

Figure 7.3: Power spectra of data $\boldsymbol{\Gamma}$, clustering signal $\boldsymbol{\kappa}$, shot noise $\boldsymbol{\zeta}$, and zero-point noise $\boldsymbol{\eta}$ when $\sigma_m = 0.01$ for R7 (*left*), R11 (*middle*), and R16 (*right*). Power is measured in 33 bins spaced according to the sampling resolution Δk derived in §3.2.2. Error bars are derived empirically as 1σ spreads from 250 realizations of each. All large-scale k-modes are pictured, while those smaller than the gridbox resolution are omitted.

the expected number of evenly distributed point-pairs at a distance r scales as $\langle N \rangle \sim r^2\, dr$.

However, the correlated zero-point noise lies along a stripe, not a spherical shell. When this planar geometry intersects the shell, a ring of radius r results. Zero-point noise clustering is only permitted in this ring where the number of point pairs scales as $r\, dr$. However, the stripe also has a non-zero width which effectively increases the number of pairs to r^p where $p \gtrsim 1$.

The zero-point noise correlation function $\xi_\eta(r)$ may be approximated as

$$\xi_{noise}(r) \propto \frac{N_\eta}{\langle N \rangle} \sim \frac{r^p\, dr}{r^2\, dr} \sim r^{p-2}. \tag{7.30}$$

Translating an isotropic correlation function to a power spectrum occurs in the usual way,

$$P_\eta(k) = 4\pi \int dr\, r^2 \frac{\sin(kr)}{kr} \xi_\eta(r) \sim \frac{1}{k^{p+1}}. \tag{7.31}$$

From this analysis, we expect $P_\eta(k)$ to go as something between k^{-2} and k^{-3}. As illustrated in Figure 7.3, this is essentially what results.

To determine how well the estimated signal $\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ approximates the power spectrum of the signal $\boldsymbol{\kappa}$, we turn to Figure 7.4 where the differences between power spectra of the data and signal are reported. This is the error one can expect under the status quo. Also plotted are the differences between the power spectra of the estimated signal and the signal, which is the error one can expect after cleansing.

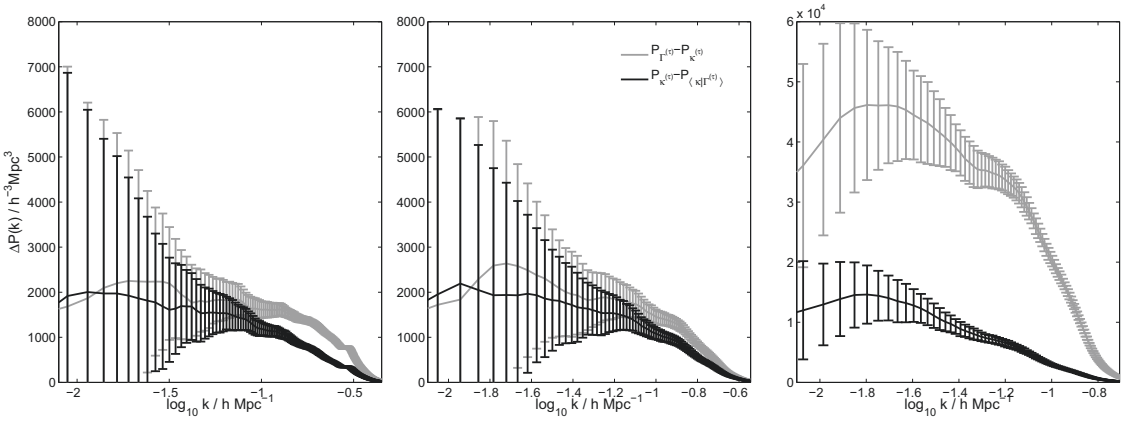

Figure 7.4: Gray lines in these panels plot the average difference between the power spectra of the raw data (i.e. signal plus noise) and the true signal for R7 (*left*), R11 (*middle*), and R16 (*right*). The black lines plot the average difference between the power spectra of the true signal and reconstructed signal. The 1σ spreads of the differences are measured using 125 realizations.

We find that for almost all k the estimated signal $\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ produce power spectra closer to the ground truth than does the uncleansed data $\boldsymbol{\Gamma}^{(\tau)}$.[1] The benefit of cleansing is most

[1]We cannot rule out the possibility that the improvement not being universal for all k is a result of an

pronounced in the R16 case where the shot noise is a greater fraction of the measured data. The improvement for R11 is modest since the overall noise variance (as reported in Table **??**) is smallest in this case.

By definition, the power of the data $P_{\boldsymbol{\Gamma}^{(\tau)}}(k)$ is always larger than that of the signal $P_{\boldsymbol{\kappa}^{(\tau)}}(k)$. In contrast we find that, on average, the power of the signal is always larger than that of the estimated signal $\hat{P}_{\langle\boldsymbol{\kappa}|\boldsymbol{\Gamma}^{(\tau)}\rangle}(k)$. This reveals that the estimated signal systematically underestimates the true power. Put another way, our cleansing process removes more noise power than is actually present.

As discussed in §3.2, the amplitudes of the clustering signal power spectrum at each wavenumber are independent of one another and uncorrelated in k-space. Our noise minimization technique, however, exploits spatial correlations in signal and noise. This introduces a form of cross-talk in which modes no longer retain their independence.

The correlation structure of $\hat{P}_{\langle\boldsymbol{\kappa}|\boldsymbol{\Gamma}^{(\tau)}\rangle}(k))$ is presented in Figure 7.5. We see that $\langle\boldsymbol{\kappa}|\boldsymbol{\Gamma}^{(\tau)}\rangle$ introduces correlations in the power that did not previously exist. The largest of these correlations are induced in adjacent k-bins. These correlations are universally positive.

In general, the correlations induced between modes is relatively weak. Over 82% of mode pairs have $|\mathrm{Corr}_{ij}| < 0.2$ for all cell sizes. Approximately 89% have $|\mathrm{Corr}_{ij}| < 0.3$. Some positive correlations are introduced on small scales for the R11 and R16 cells. In the case of R11, the positive correlations are more sporadic but extend to smaller k. The positive correlation with R16 are more uniform, but localized at the largest k values.

insufficient number of realizations.

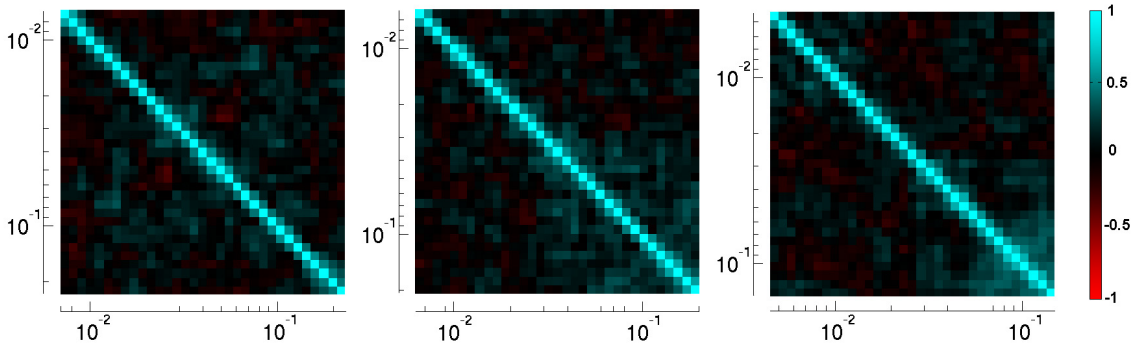

Figure 7.5: Correlations between band powers of the estimated signal's power spectrum for R7 (*left*), R11 (*middle*), and R16 (*right*). The correlations $\mathrm{Corr}_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$, where $C_{ij} \equiv \mathrm{Cov}\left(\hat{P}_{\langle\boldsymbol{\kappa}|\boldsymbol{I}^{(\tau)}\rangle}(k_i)\,,\hat{P}_{\langle\boldsymbol{\kappa}|\boldsymbol{I}^{(\tau)}\rangle}(k_j)\right)$ are depicted using a red/black/blue color scale. Each image is a 33×33 pixel symmetric matrix where the k_i and k_j of the numerical correlation coefficients are indicated by the vertical and horizontal scales. These scales are in units of $h^{-1}\mathrm{Mpc}$.

7.5.3 Metropolis-Hastings Verification

In this section we provide details about the implementation of the Metropolis-Hastings algorithm to empirically verify the cleansing result obtained analytically. We limit the focus of this section to independently verifying that $\hat{\boldsymbol{\kappa}}$ converges to $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$.

One of the trickier parts of running the sampler is calculating β since $q(\theta) \approx 10^9$, if one ignores the exponent. When taken to the power $-(N+f)/2$, the value reduces to zero (i.e. beyond machine precision). As a result, we need to be clever during the calculation to avoid numerical overruns. We found it effective to work with the exponentials of natural logs,

$$\begin{aligned}\frac{g(\boldsymbol{\theta}'|\boldsymbol{\delta})q(\boldsymbol{\theta})}{g(\boldsymbol{\theta}|\boldsymbol{\delta})q(\boldsymbol{\theta}')} &= \frac{\exp\left(\ln\left(g(\boldsymbol{\theta}'|\boldsymbol{\delta})\right)\right)}{\exp\left(\ln\left(g(\boldsymbol{\theta}|\boldsymbol{\delta})\right)\right)}\frac{\exp\left(\ln\left(q(\boldsymbol{\theta})\right)\right)}{\exp\left(\ln\left(q(\boldsymbol{\theta}')\right)\right)} \\ &= \exp\Bigg(l(\boldsymbol{\theta}'|\boldsymbol{\delta}) - l(\boldsymbol{\theta}|\boldsymbol{\delta}) \\ &\quad + \left(\frac{N+f}{2}\right)\left(\ln\left(a(\boldsymbol{\theta}')\right) - \ln\left(a(\boldsymbol{\theta})\right)\right)\Bigg) \qquad (7.32)\end{aligned}$$

where $a(\boldsymbol{\theta}) \equiv 1 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\Lambda}^{(W)} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$.

This facilitates the next step—adjusting the t-distribution degrees of freedom parameter f to admit the optimal number of variates. The ideal acceptance rate for an N-dimensional Gaussian has been shown to be about 23% when N samples are drawn. Because different diagnostic test vectors $\boldsymbol{\delta}$ only shift $\hat{\boldsymbol{\theta}}$ and not the curvature $\mathbf{Y}$, the results in Figure 7.6 are independent of one's choice of $\boldsymbol{\delta}$.

The progress of a Metropolis-Hastings algorithm is assessed using a "trace plot" whereby each time a new $\boldsymbol{\theta}^{(\tau)}|\boldsymbol{\delta}$ is drawn, the value of its i^{th} element is plotted. If the initial vector $\boldsymbol{\theta}^{(0)}$ is selected far from the peak of $g(\boldsymbol{\theta}|\boldsymbol{\delta})$, random vectors $\boldsymbol{\theta}'$ must be drawn until a region containing the higher probabilities is reached. This is known as the "burn-in" period. The number of vectors needed, and ultimately discarded, during the burn-in period varies with distribution and initial position. On a stable trace plot, i.e. one from which we can accept random variates, the vector elements drawn will vary around a fixed horizontal trend line. A burn-in period typically manifests itself on a trace plot as a trend approaching that baseline.

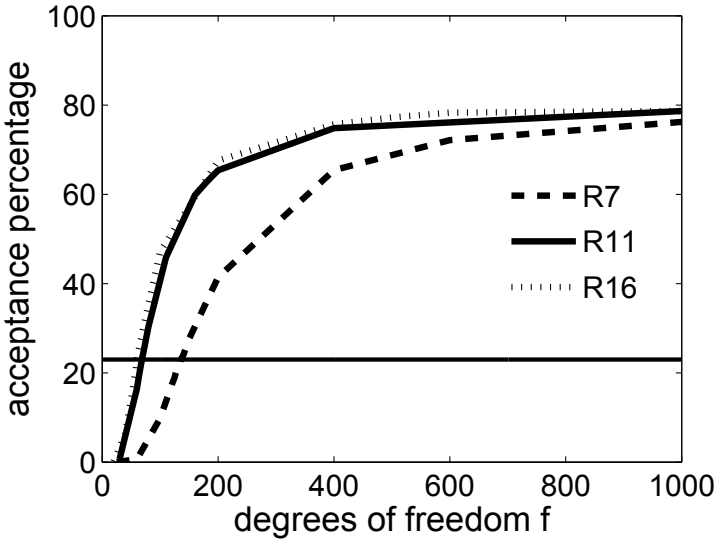

Figure 7.6: Percentage of variates accepted through the Metropolis-Hastings algorithm as a function of the degrees of freedom parameter f. The horizontal line indicates the ideal acceptance percentage of 23%, which is reached at $f \approxeq 136$ for R7, $f \approxeq 69$ for R11, and $f \approxeq 60$ for R16.

We know precisely where $g(\boldsymbol{\theta}|\boldsymbol{\delta})$ peaks, so no burn-in period is necessary. We verify this assumption by examining in Figure 7.7 trace plots of the first 1000 accepted draws along four dimensions selected to lie at well-separated redshifts. We see the trace plots are stable from the outset, meaning we can start accepting $\boldsymbol{\theta}^{(\tau)}|\boldsymbol{\delta}$ right away. Note that the variates $\boldsymbol{\kappa}^{(\tau)}|\boldsymbol{\delta}$ tend to cluster more around the signal than the data even though those variates were generated *without explicit knowledge of* $\boldsymbol{\delta}_\kappa$. We find that regardless of the dimensionality of our problem, roughly 10^5 realizations are necessary before the signal estimate converges. This result appears to be uniform along all dimensions of the recovered signal as shown in Figure 7.8.

Using 10^5 signal realizations, we test the consistency of the Metropolis-Hastings estimate $\hat{\boldsymbol{\kappa}}$ against $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$ by taking the 2-norm difference between the two solutions. Table

Radius (h^{-1}Mpc)	$\|\|\hat{\boldsymbol{\kappa}} - \langle\boldsymbol{\kappa}\|\boldsymbol{\delta}\rangle\|\|_2$
7	0.473 ± 0.002
11	0.2189 ± 0.0011
16	0.1496 ± 0.0012

Table 7.2: Distance between the analytic signal solution $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$ and empirically-derived Metropolis-Hastings solution $\hat{\boldsymbol{\kappa}}$ as quantified through the 2-norm.

7.2 summarizes the results. For each cell size, the distance between the two solutions is less than 1% of the distance between $\boldsymbol{\delta}$ and $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$. Equation (7.11) is therefore strongly supported by the empirical approach of the Metropolis-Hastings algorithm.

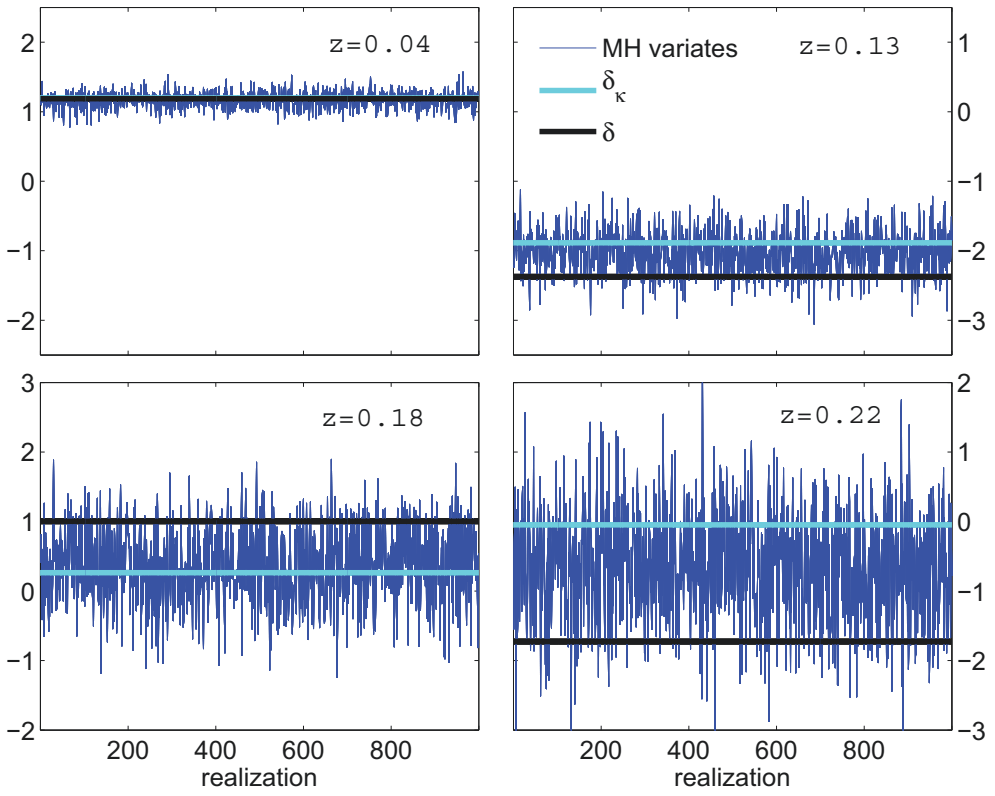

Figure 7.7: Trace plots of four randomly selected R7 elements at well-separated redshifts. The dark blue line follows the variate in that dimension from one accepted realization to the next. The black and cyan lines indicate respectively the values of δ_i and $\delta_{\kappa,i}$ for that dimension. Variates have a greater variance at higher redshifts where the shot noise and zero-point noise have the greatest impact.

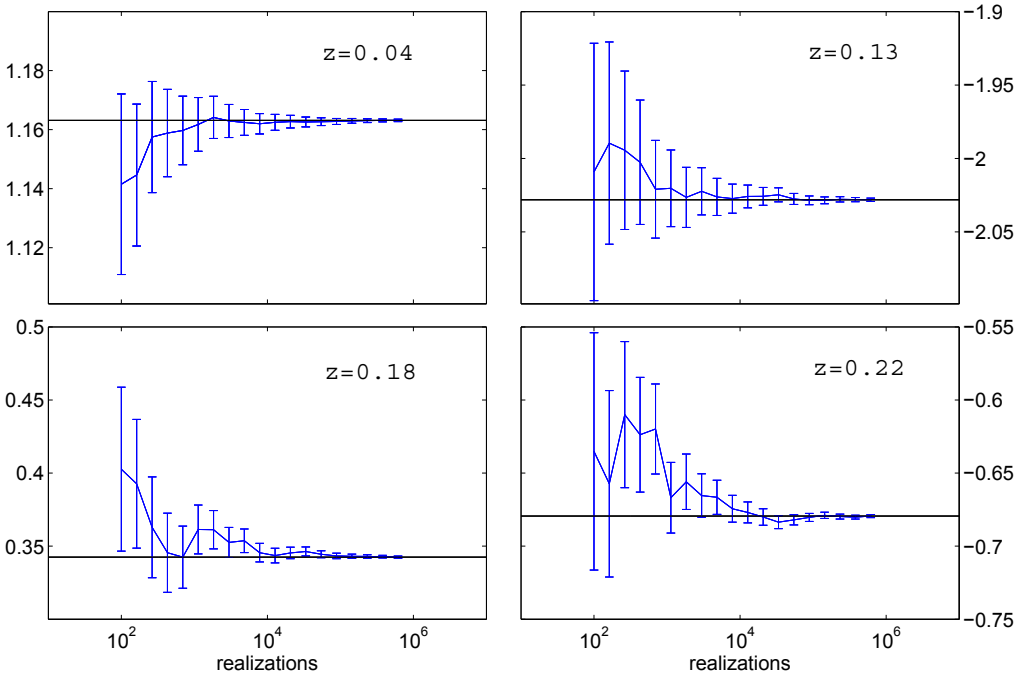

Figure 7.8: Average value of Metropolis-Hastings random variates as a function of the number of realizations. Error bars are one standard deviation of the estimated error on the mean, $\sqrt{\sigma^2_{\kappa_i}/K}$, where K is the number of realizations. This error formula is merely an approximation since random variates drawn through Metropolis-Hastings are technically not independent. However, they are drawn from an independent candidate density that blankets the entire distribution $g(\boldsymbol{\theta}|\boldsymbol{\delta})$, making essentially all of the parameter space accessible on each draw. Consequently, the correlations should be relatively weak.

7.6 Noise Estimation

In this section we estimate both the shot noise and systematic noise from a set of data realizations $\boldsymbol{\Gamma}^{(\tau)}$. We analyze our results in two ways. First, we attempt to recover a fixed shot noise vector $\boldsymbol{\zeta}^{(0)}$ that has been added to random realizations of signal plus systematic noise. Second, we allow all three components to vary and investigate whether using our estimator is better than naively assuming that the shot noise in each cell takes its most likely value of zero. In §7.6.2 we employ a similar analysis for the systematic noise, and introduce a prediction of $\Delta\boldsymbol{m}$ as well.

7.6.1 Shot Noise

Our first test of the shot noise estimator is an attempt to recover a fixed, randomly selected overdensity vector $\boldsymbol{\zeta}^{(0)}$. We construct data vectors $\boldsymbol{\Gamma}^{(\tau)}$ using K signal and zero-point realizations such that $\boldsymbol{\Gamma}^{(\tau)} = \boldsymbol{\kappa}^{(\tau)} + \boldsymbol{\eta}^{(\tau)} + \boldsymbol{\zeta}^{(0)}$. Using equation (7.17), we estimate the shot noise $\langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle$ for each cell i and average the estimates,

$$\hat{\zeta}_i^{(0)} = \frac{\sum_{\tau=1}^{K} \langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle}{K}. \tag{7.33}$$

This test is performed for each cell size and the results are plotted in Figures 7.9, 7.10, and 7.11. The R7 results illustrated in Figure 7.9 indicate that our shot noise estimator does reasonably well in determining whether a particular shot noise component is positive or negative. The roughly linear trend indicates that the signal estimates per cell are relatively

consistent, i.e. that the ratio $\zeta_i^{(0)}/\hat{\zeta}_i^{(0)}$ is fixed plus or minus some random scatter. This relationship is relaxed for the highest redshift cells which tend to "curl back" to the unit slope, suggesting that the estimator improves with the magnitude of the shot noise. The slope of the solution space is less than 1, revealing that shot noise estimates on a per-cell basis are likely to be conservative.

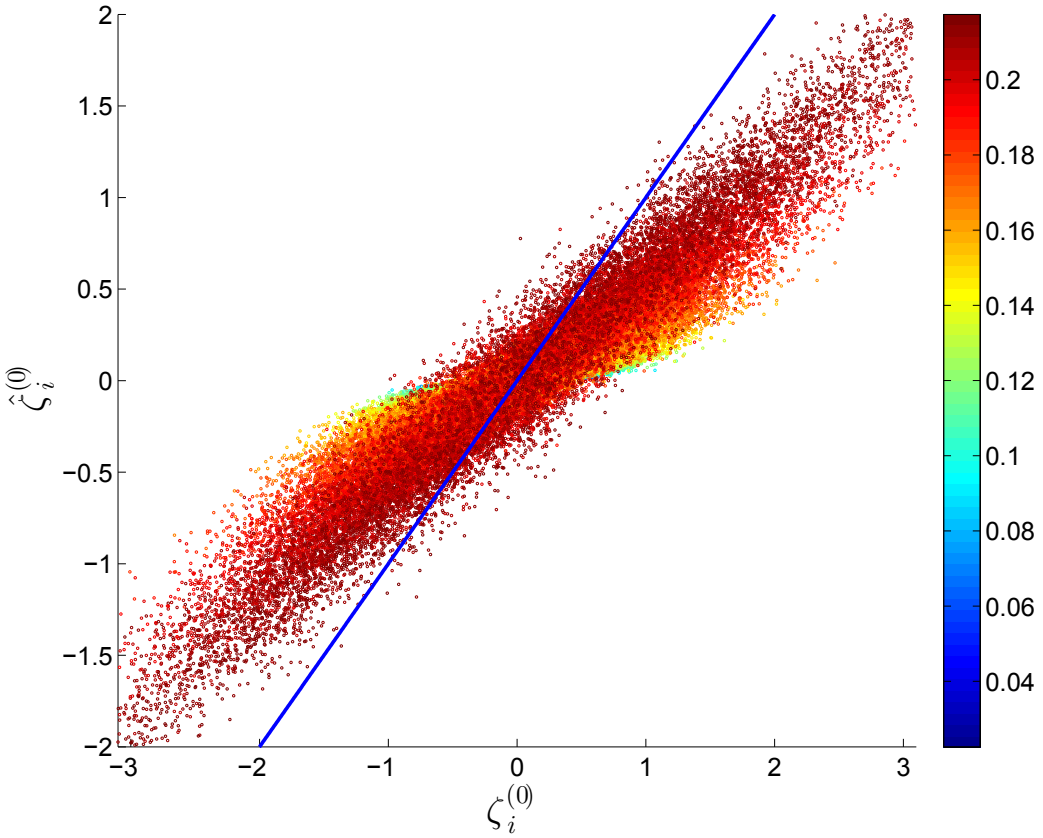

Figure 7.9: Attempted recovery of a single shot noise realization $\boldsymbol{\zeta}^{(0)}$. Each point represents one R7 cell where color marks its redshift. The horizontal axis marks the fixed value of the shot noise overdensity in each cell. The vertical axis reports the average of the $\langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle$ solutions in each cell using 10,000 realizations of signal plus zero-point systematic noise. The blue line has unit slope. Points along this line have estimated shot noise values that exactly match their true values.

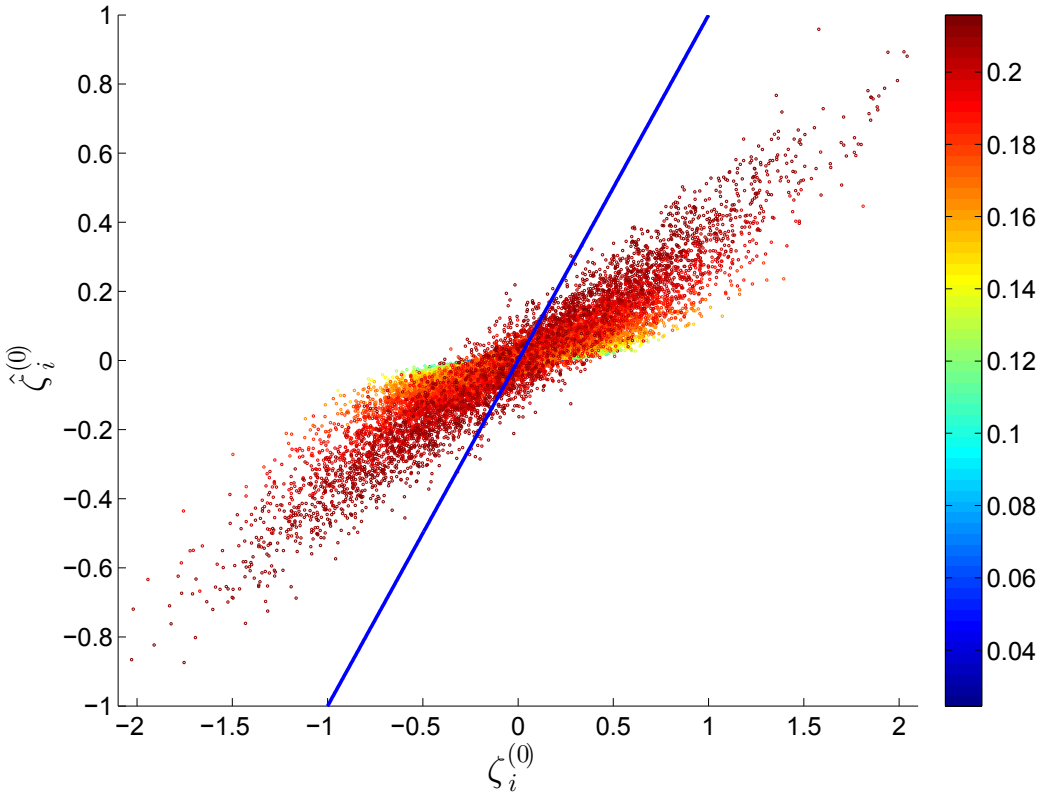

Figure 7.10: Same as Figure 7.9 but for R11 cells.

The results of the R16 test (see Figure 7.11) reveal a sort of "redshift dispersion" whereby the efficacy of the estimator depends even more strongly upon the redshift of the cell. The lowest redshift cells (i.e. $z < 0.2$) have the least shot noise on average. Here, the predictive power of equation (7.17) is limited. However, as the redshift of a cell increases, so too does the estimator's ability to recover $\zeta_i^{(0)}$. For cells at $z \approx 0.3$, this recovery is almost complete. Regardless of redshift, the results are almost always in the correct quadrant, that is, $\hat{\zeta}_i^{(0)}$ has the same sign as $\zeta_i^{(0)}$.

Our second test of the shot noise estimator allows all three data components, including

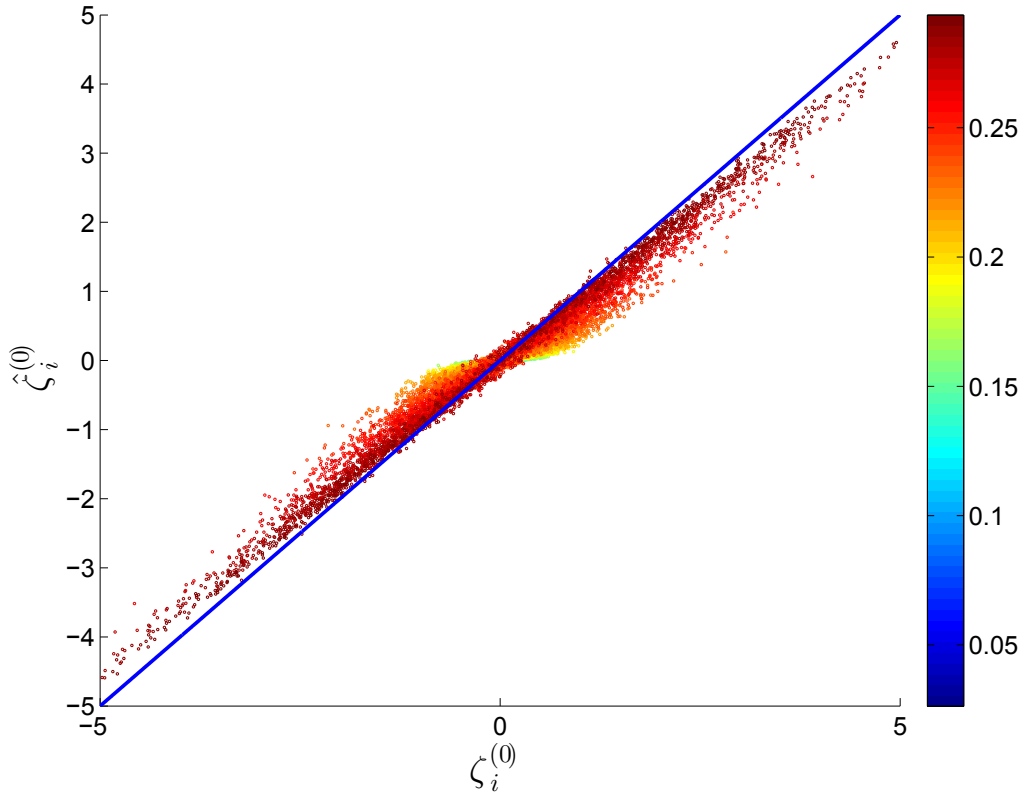

Figure 7.11: Same as Figure 7.9 but for R16 cells.

the shot noise, to vary such that $\boldsymbol{\Gamma}^{(\tau)} = \boldsymbol{\kappa}^{(\tau)} + \boldsymbol{\eta}^{(\tau)} + \boldsymbol{\zeta}^{(\tau)}$. We employ this test so as to avoid drawing too strong a conclusion from a single shot noise vector $\boldsymbol{\zeta}^{(0)}$.

We compare our method to the null hypothesis that the shot noise in each cell takes its most likely value, zero. We quantify the performance of both approaches using the difference between the truth and either zero or my estimate. To this end we define the *default error* for cell i during realization τ to be $|\zeta_i^{(\tau)}|$. We compare this against our *estimate error* $|\zeta_i^{(\tau)} - \langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle|$. For each cell, we average these errors over all realizations to see which approach yields the better estimate.

Figure 7.12 shows that $\langle \boldsymbol{\zeta} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ does a better job of estimating the shot noise than does a default guess of zero. For low redshift cells, there is little statistical difference between assuming the shot noise is zero and assuming it is $\langle \boldsymbol{\zeta} | \boldsymbol{\Gamma}^{(\tau)} \rangle$. This is expected since the shot noise component in low redshift cells is very nearly zero and both approaches reflect as much.

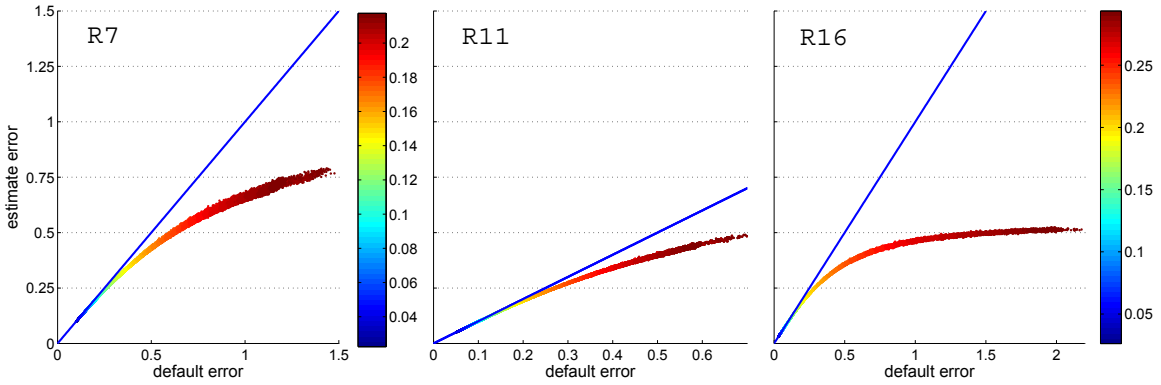

Figure 7.12: A comparison between estimating shot noise using my method versus assuming a shot noise of zero. Each point represents a single cell where its color denotes redshift. The left colorbar represents the R7 and R11 cases, while the right colorbar represents the R16 case. The horizontal axis captures $\langle |\zeta_i^{(\tau)}| \rangle$, the average shot noise error in each cell. The vertical axis quantifies $\langle |\zeta_i^{(\tau)} - \langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle| \rangle$, the average error in the shot noise when the default guess of $\zeta_i = 0$ is replaced with $\langle \zeta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle| \rangle$. The blue line has unit slope. Cells along this line display no difference between their default error and estimate error. Averages are taken over 10,000 realizations.

The benefit of my method is clearer at high redshifts where shot noise is of greater magnitude. Cells trend away from the unit line, which indicates that the estimate error grows more slowly than the default error. Note that this result holds without placing any constraints upon the signal, shot noise or systematic noise beyond what is conveyed through their respective covariance functions.

7.6.2 Systematic Noise

We repeat the analysis of §7.6.1 for our systematic zero-point errors. First, we generate a single zero-point realization $\Delta\boldsymbol{m}^{(0)}$ and an associated $\boldsymbol{\eta}^{(0)} = \mathbf{A} \cdot \Delta\boldsymbol{m}^{(0)}$. We blend this into K data realizations with randomized signal and shot noise vectors $\boldsymbol{\Gamma}^{(\tau)} = \boldsymbol{\kappa}^{(\tau)} + \boldsymbol{\eta}^{(0)} + \boldsymbol{\zeta}^{(\tau)}$. Each realization admits a solution for the expected signal and shot noise, and thus for the zero-point noise as well,

$$\langle \boldsymbol{\eta} | \boldsymbol{\Gamma}^{(\tau)} \rangle = \boldsymbol{\Gamma}^{(\tau)} - \langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle - \langle \boldsymbol{\zeta} | \boldsymbol{\Gamma}^{(\tau)} \rangle. \tag{7.34}$$

The zero-point solutions are averaged to provide a best estimate in each cell,

$$\hat{\eta}_i^{(0)} = \frac{\sum_{\tau=1}^{K} \langle \eta_i | \boldsymbol{\Gamma}^{(\tau)} \rangle}{K}. \tag{7.35}$$

Comparisons between $\boldsymbol{\eta}^{(0)}$ and $\hat{\boldsymbol{\eta}}^{(0)}$ are provided in Figures 7.13, 7.14, 7.15. As with the shot noise, the R7 estimator does an adequate job of estimating the signs of the zero-point overdensities, and the correlation between the truth and the estimate is roughly linear. In contrast, however, the estimates are about an order of magnitude smaller than the true values and seem to be less tightly coupled to redshift.

We encountered less success estimating the zero-point noise for the R11 and R16 cells, however. The symmetric, circular distribution of points in Figures 7.14 and 7.15 reveal that equation (7.19) has little predictive power when the number of PRIMARY SEGMENTs projected onto each cell is large. Furthermore, the noise introduced when $\sigma_m = 0.01$ is

less than a 1% effect, meaning that the amount of information we are trying to extract is very small. The failures of the R11 and R16 cases are not an indictment of the method as a whole, but rather an expression of its limitations.

To examine the role played by the photometric zero-points themselves, we use $\hat{\boldsymbol{\eta}}^{(0)}$ to compute a best-fit set of photometric coefficients $\Delta\hat{\boldsymbol{m}}^{(0)}$ by solving the overdetermined linear equation,

$$\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{A} \cdot \Delta\hat{\boldsymbol{m}}^{(0)}. \tag{7.36}$$

These results, which we present in Figure 7.16, are more edifying. Like $\hat{\boldsymbol{\eta}}^{(0)}$, the estimates for the zero-points are very conservative. But from this perspective we see that the predictive power of $\Delta\hat{\boldsymbol{m}}^{(0)}$ is related to the lengths of the PRIMARY SEGMENTs themselves. Because short PRIMARY SEGMENTs intersect relatively few cells, there is little available information to constrain their values and estimates of $\Delta\hat{m}_i^{(0)} \approx 0$ are returned. But zero-point estimates for longer PRIMARY SEGMENTs (those with lengths $\gtrsim 40^\circ$) tend to depart from the line of slope zero and approach their true values.

These observations help explain the "linear offshoot" features in Figure 7.13. In the upper right-hand corner of the figure a line of cells shoots out at a steeper slope that more closely matches the truth. Figure 7.17, which provides a visual representation of the locations of those cells, illustrates that they all intersect the longest PRIMARY SEGMENTs in the DR6 survey. This supports the idea that the more cells a particular systematic effect affects, the better its estimate will be.

To assess the systematic noise estimator for an ensemble of zero-point simulations, we use multiple realizations of $\Delta\boldsymbol{m}^{(\tau)}$ and $\boldsymbol{\eta}^{(\tau)} = \mathbf{A} \cdot \Delta\boldsymbol{m}^{(\tau)}$ to calculate $\langle\boldsymbol{\eta}|\boldsymbol{\Gamma}^{(\tau)}\rangle$ and $\langle\Delta\boldsymbol{m}|\boldsymbol{\Gamma}^{(\tau)}\rangle$ for each $\boldsymbol{\Gamma}^{(\tau)} = \boldsymbol{\kappa}^{(\tau)} + \boldsymbol{\eta}^{(\tau)} + \boldsymbol{\zeta}^{(\tau)}$. As with the shot noise, the *default error* under the null hypothesis that $\eta_i = 0$ is $|\eta_i^{(\tau)}|$ while the *estimate error* is $|\eta_i^{(\tau)} - \langle\eta_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|$. Figure 7.18 compares the two in the R7 case.

For nearly all cells the default error is larger than the estimate error, indicating that it is better to approximate the zero-point overdensities using equation (7.19) than to assume $\boldsymbol{\eta} = 0$. The trend is that as the default error increases (largely as a function of redshift), so too does the benefit of using our estimator. This benefit is most pronounced for the cells along DR6's longest PRIMARY SEGMENTs. These cells comprise the thin strip that runs through the figure's central diagonal. Moreover, our zero-point overdensity estimator is robust in that only a very small number of cells are not benefited by the process. The majority of these are the highest redshifts cells, reflecting a similar feature seen in Figure 7.12.

Finally, we repeat this analysis for the photometric zero-points. We assume a default guess of $\Delta m_i = 0$ for the zero-point in each PRIMARY SEGMENT and define the *default error* as $|\Delta m_i^{(\tau)}|$ and the *estimate error* as $|\Delta m_i^{(\tau)} - \langle\Delta m_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|$.

The averages of these errors are shown in Figure 7.19. While there is little-to-no improvement for the smallest PRIMARY SEGMENTs, our method offers a better prediction than the default for longer PRIMARY SEGMENTs. The greatest gains are experienced by the longest PRIMARY SEGMENTs.

These results highlight the challenge present in directly predicting small, systematic errors—they can very easily become obscured by the much larger signal and shot noise components. This is made even more difficult when limited intersections between cells and PRIMARY SEGMENTs do not provide enough information to make reliable statistical inferences.

That said, it is encouraging that these results trend in the right direction for the R7 case where cells are small, numerous and local. The signs of the errors are adequately well-predicted as are their relative magnitudes. And while the gains are modest, statistically one is better off using these error estimates rather than naively assuming they equal zero.

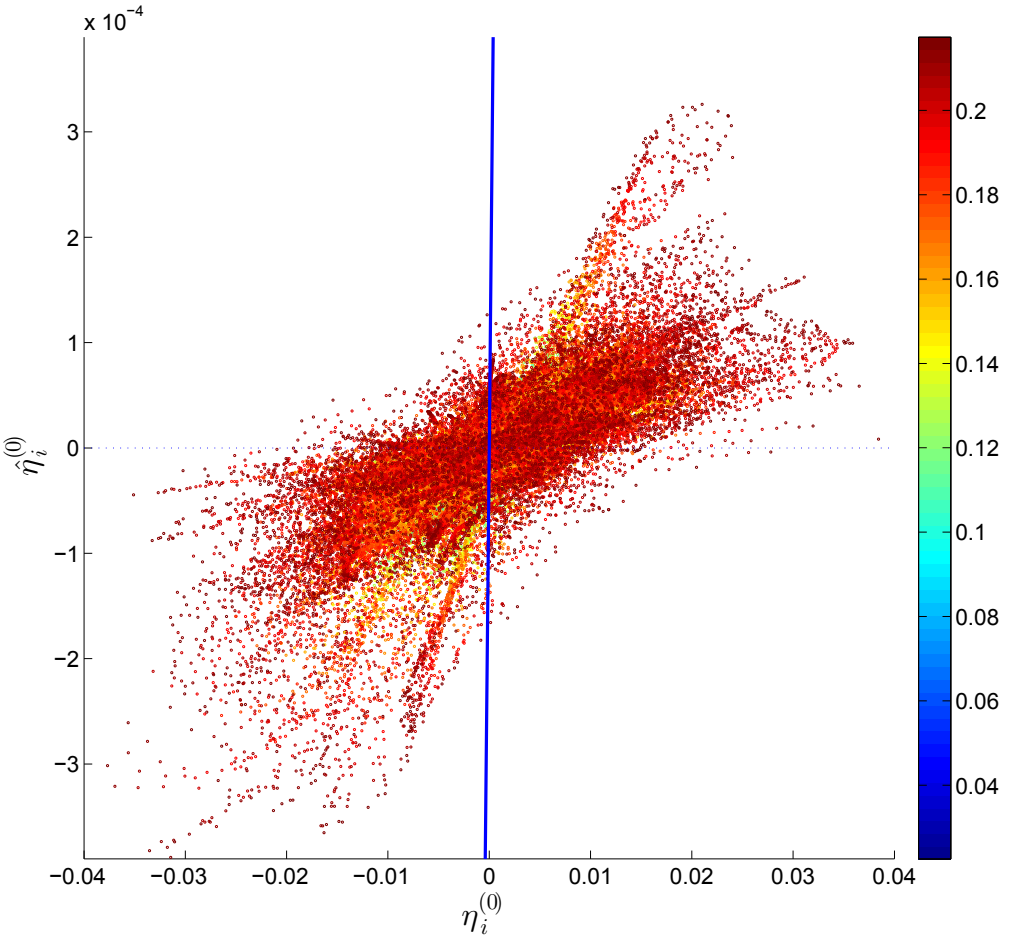

Figure 7.13: Attempted recovery of a single systematic noise realization $\boldsymbol{\eta}^{(0)}$. Each point represents one R7 cell where color marks its redshift. The horizontal axis marks the fixed value of the zero-point overdensities in each cell. The vertical axis reports the average of the $\langle \boldsymbol{\eta} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ solutions in each cell using 10,000 realizations of signal plus shot noise. The blue line has unit slope. Points along this line have estimated systematic noise values that exactly match their true values.

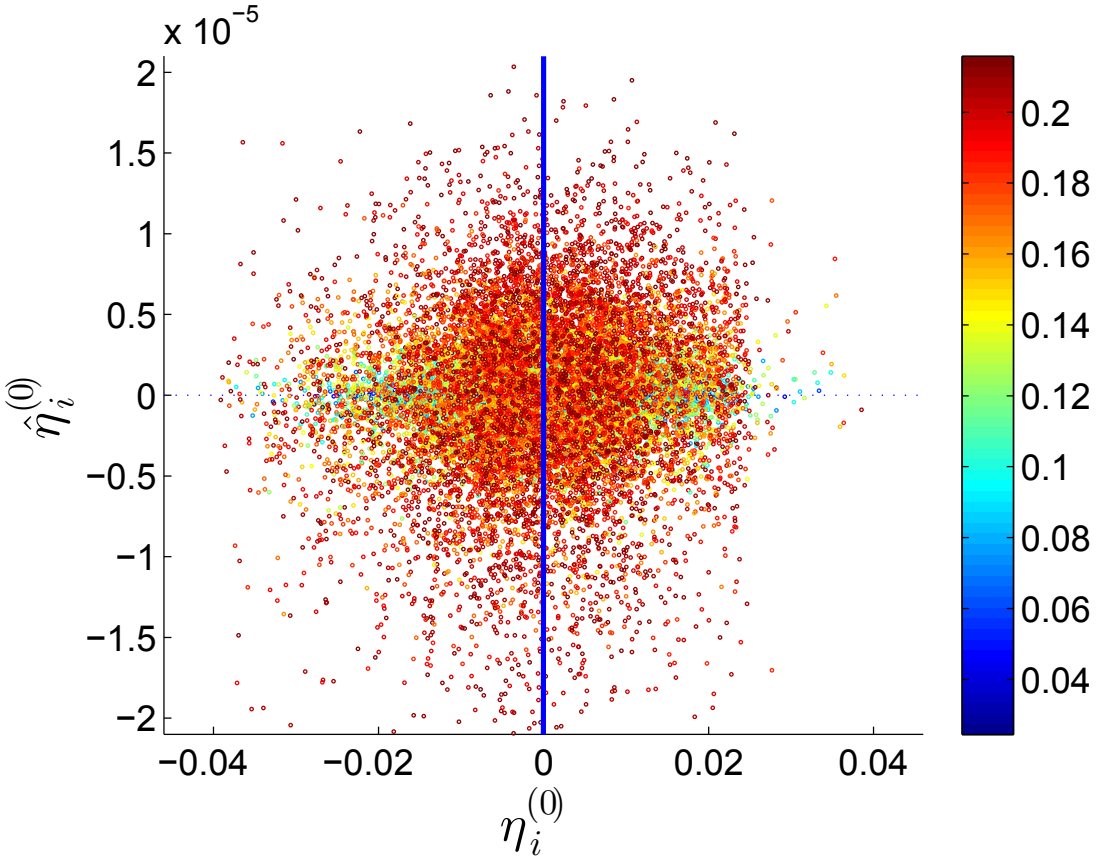

Figure 7.14: Same as Figure 7.13 but R11 cells.

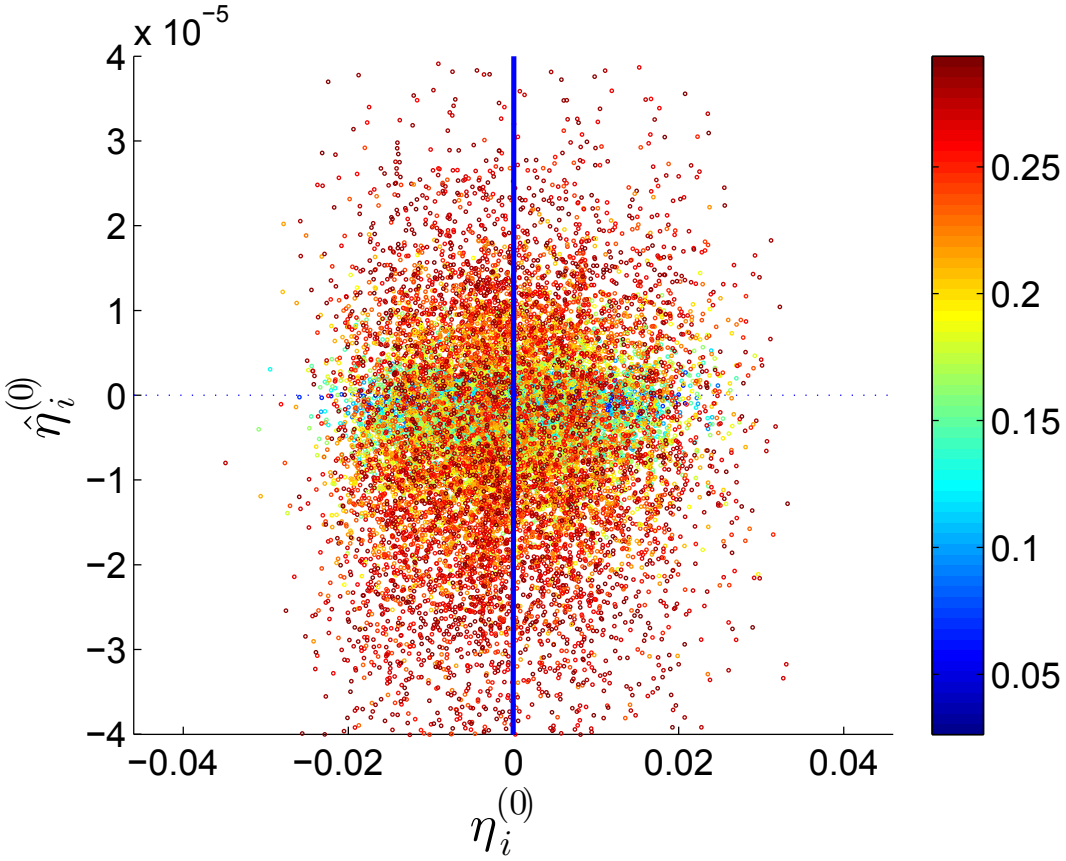

Figure 7.15: Same as Figure 7.13 but R16 cells.

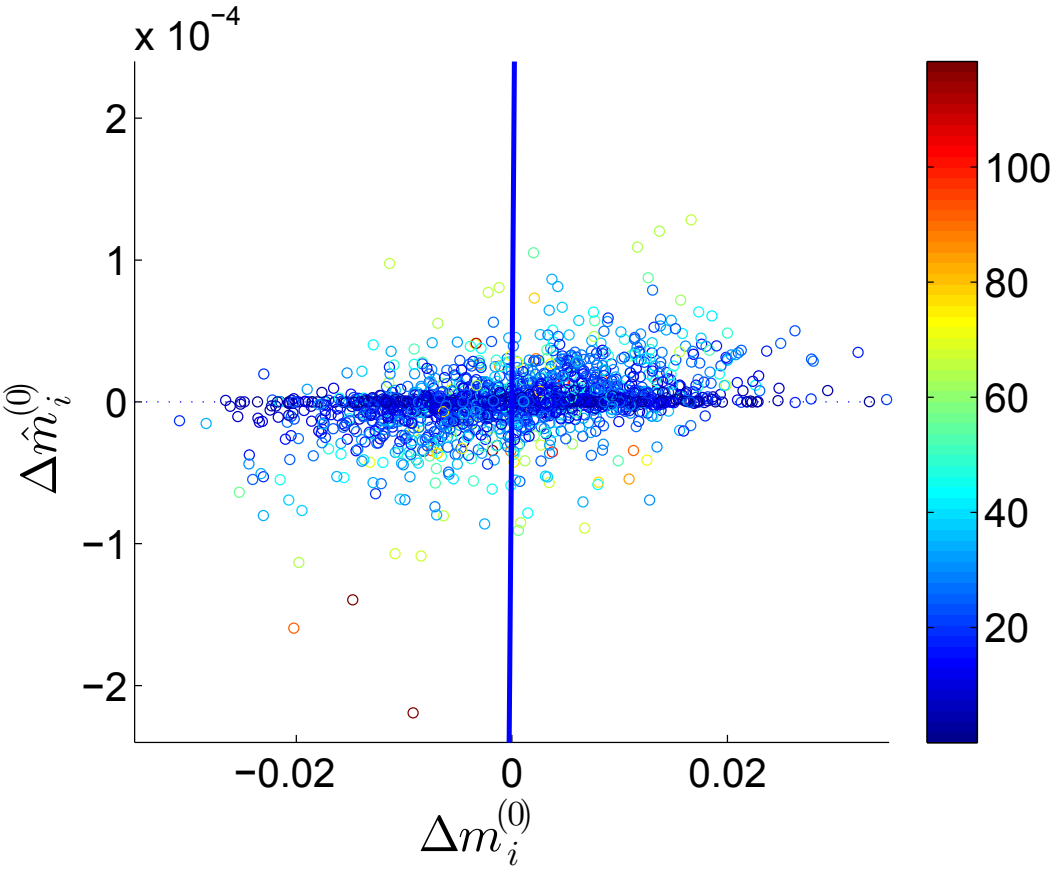

Figure 7.16: Relationship between the true and estimated zero-point calibration offsets. Each point represents an individual PRIMARY SEGMENT where color marks its length in degrees. The horizontal axis plots the true photometric offset while the vertical axis plots its estimate from equation (7.36). The blue line has unit slope where points along it have perfectly predicted offsets.

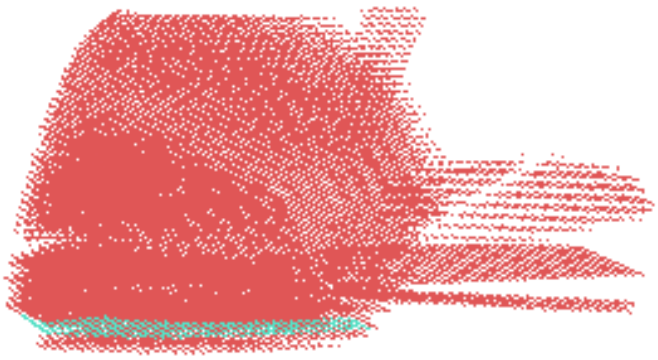

Figure 7.17: Locations of the R7 cells that comprise the linear offshoot feature in the upper right quadrant of Figure 7.13. Red dots mark the locations of cells in the three-dimensional DR6 spectroscopic footprint. Cells in cyan are those that lie within the line feature.

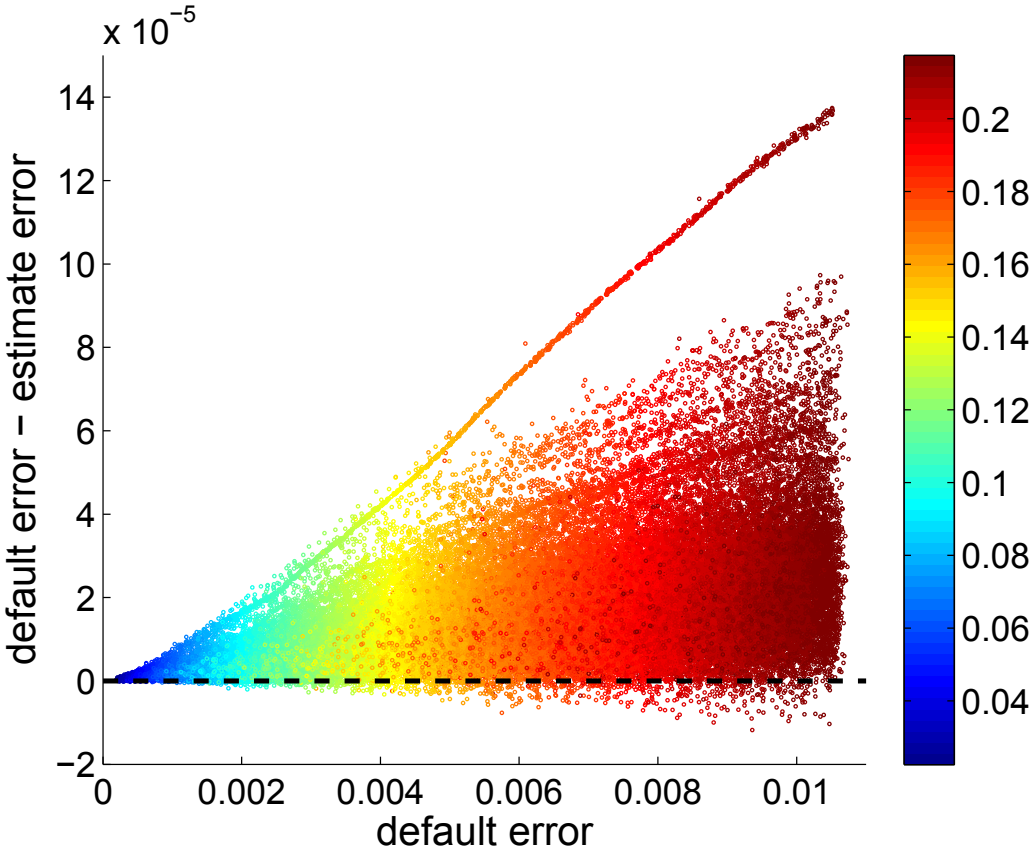

Figure 7.18: A comparison between estimating zero-point noise using our method versus assuming zero-point overdensities of zero in the R7 case. This figure is similar in structure to Figure 7.12, except it replaces $\langle|\zeta_i^{(\tau)}|\rangle$ and $\langle|\zeta_i^{(\tau)} - \langle\zeta_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|\rangle$ on the horizontal and vertical axes with $\langle|\eta_i^{(\tau)}|\rangle$ and $\langle|\eta_i^{(\tau)} - \langle\eta_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|\rangle$ respectively. The black dotted line takes the place of the unit slope in Figure 7.12.

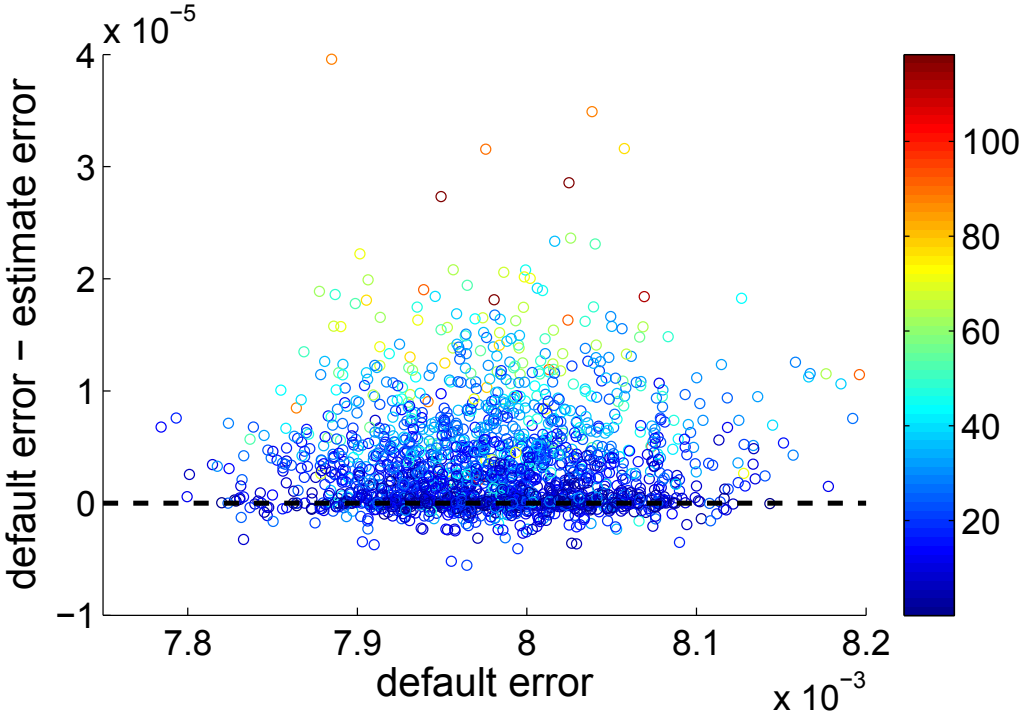

Figure 7.19: A comparison between estimating the photometric zero-points using our method versus assuming the zero-points equal zero in the R7 case. This figure is identical in structure to Figure 7.18, except this replaces $\langle|\eta_i^{(\tau)}|\rangle$ and $\langle|\eta_i^{(\tau)} - \langle\eta_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|\rangle$ on the horizontal and vertical axes with $\langle|\Delta m_i^{(\tau)}|\rangle$ and $\langle|\Delta m_i^{(\tau)} - \langle\Delta m_i|\boldsymbol{\Gamma}^{(\tau)}\rangle|\rangle$ respectively. Color marks the length of the SEGMENTs in degrees.

Chapter 8

Data Cleansing – Application

Throughout this dissertation we have examined ways to minimize errors and uncertainties in galaxy counts and overdensities. In Chapter 5, we improved the geometric descriptions of the SDSS photometric and spectroscopic footprints. This enabled us to study regions with relatively uniform completeness properties, and better quantify the expected numbers of galaxies in cells.

In Chapter 6, we compared methods to count MGS objects in cells. We demonstrated that the optimal counting strategy depends upon the region type and the cell's size and redshift. The final product was a census of δ, the overdensities of DR6 MGS targets in cells. In Chapter 7 we introduced a method that can reduce shot and systematic noise, if provided reasonably accurate signal and noise models to function as Bayesian priors.

In this chapter, we incorporate all of these tools to deliver an optimal account of MGS targets in cells. Data gathered during Chapters 5 and 6 are processed with the cleansing

method of Chapter 7. We report how this cleansing affects the distributions of cells' galaxy counts and overdensities. We show how the MGS power spectra and one-dimensional 2PCFs are modified. We conclude with a test that determines whether the cleansing process is removing noise, or simply reducing spectral power in general.

8.1 Cell Statistics

Our process to reduce noise operates within a Bayesian framework. Prior information is encoded in mean-zero signal and noise models. Interpreted at its most basic level, these Bayesian priors suggest that if $\delta_i > 0$, the error due to noise is most likely positive. Likewise, negative overdensities are more likely to contain noise that underestimates the number of galaxies present.

Therefore, the effect of cleansing should be to "contract" the overdensity histogram — to drive overdensities back towards zero. Figures 8.1 and 8.2 demonstrate the effect this has on $\boldsymbol{\delta}$. For all cell sizes, the numbers of highly overdense cells are reduced. This reduction is more pronounced for R7 and R16 due to their larger shot noise components. The modest shift observed in the R11 cells mirrors the simulation results reported in Chapter 7.

Before cleansing, the percentages of empty cells (i.e. $\delta = -1$) for R7, R11, and R16 were 43%, 14%, and 30% respectively. During cleansing, each cell's overdensity changes by a nonzero amount. Cells that were previously empty might now "contain galaxies", at least mathematically. Figure 8.3 shows that cleansing adjusted the distribution of nega-

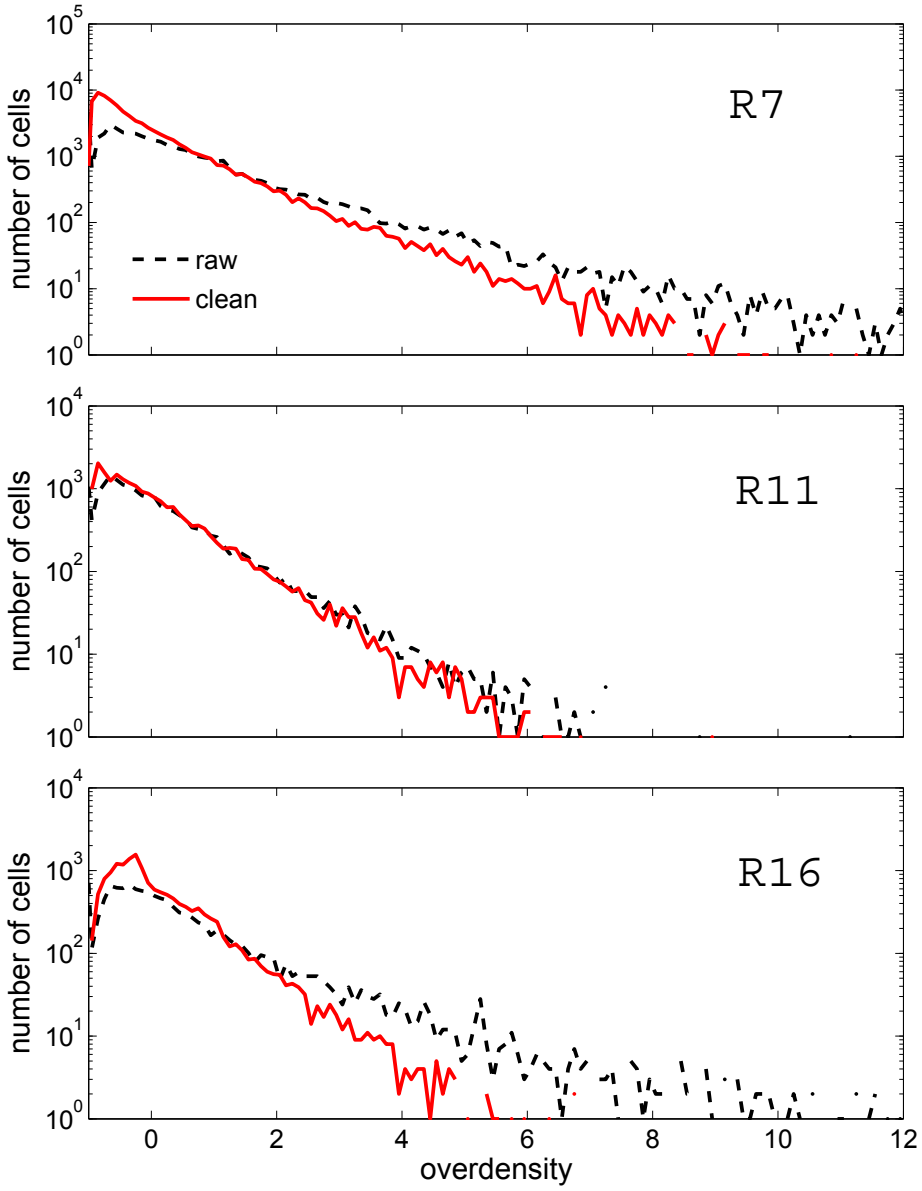

Figure 8.1: Histogram of MGS overdensities before and after cleansing. The black dashed curve replicates the results of Figure 6.36 by reporting the overdensities after accounting for MGS objects. The red curve traces the distribution of overdensities after minimizing shot noise and zero-point noise through equation (7.11). Galaxies are counted in bins of width $\Delta\delta = 0.1$.

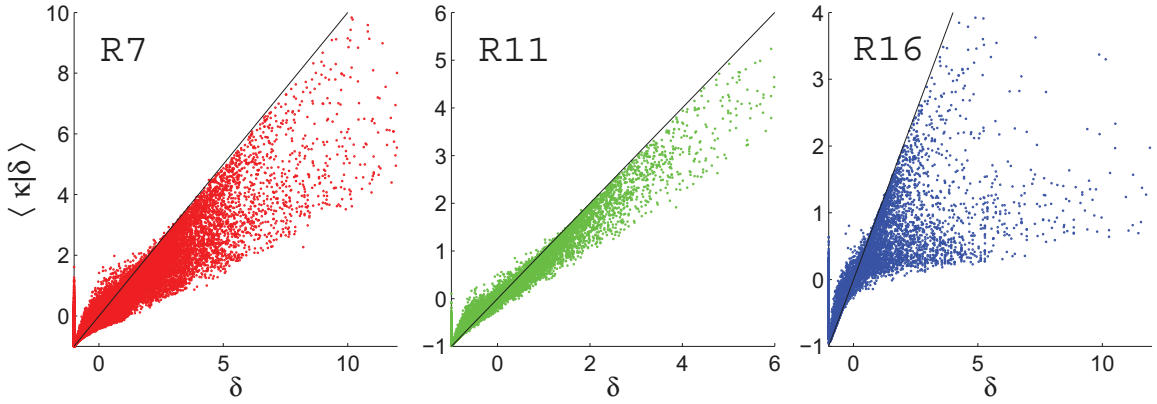

Figure 8.2: Overdensities in cells before and after cleansing. Each pixel represents a single cell. The unit slope is marked in black for clarity.

tive overdensities dramatically. The "spike" at $\delta = -1$ transformed into a more gradual distribution for $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$.

It is mathematically possible for the cleansing process to yield $\langle \kappa_i | \boldsymbol{\delta} \rangle < -1$. This result would imply that the cell contains a negative number of galaxies, and is clearly nonphysical. In our case, such instances were rare. Among R7 cells, only 76 were shifted into this range and of those, the average estimated overdensity was -1.006. None of the R11 or R16 cells experienced a change of this kind.

The overdensity shifts $\langle \kappa_i | \boldsymbol{\delta} \rangle - \delta_i$ for each cell vary as a function of redshift. Figure 8.4 reveals minimal overdensity adjustments at low redshifts, where both shot and systematic noise are at their smallest. By extension, at a fixed redshift smaller cells are shown to experience larger noise corrections on average. The range of corrections skews negative since, unlike negative overdensities, the upper limit of δ is unbounded. The cleansing algorithm identifies highly overdense cells as more likely to contain target count-increasing sources of noise.

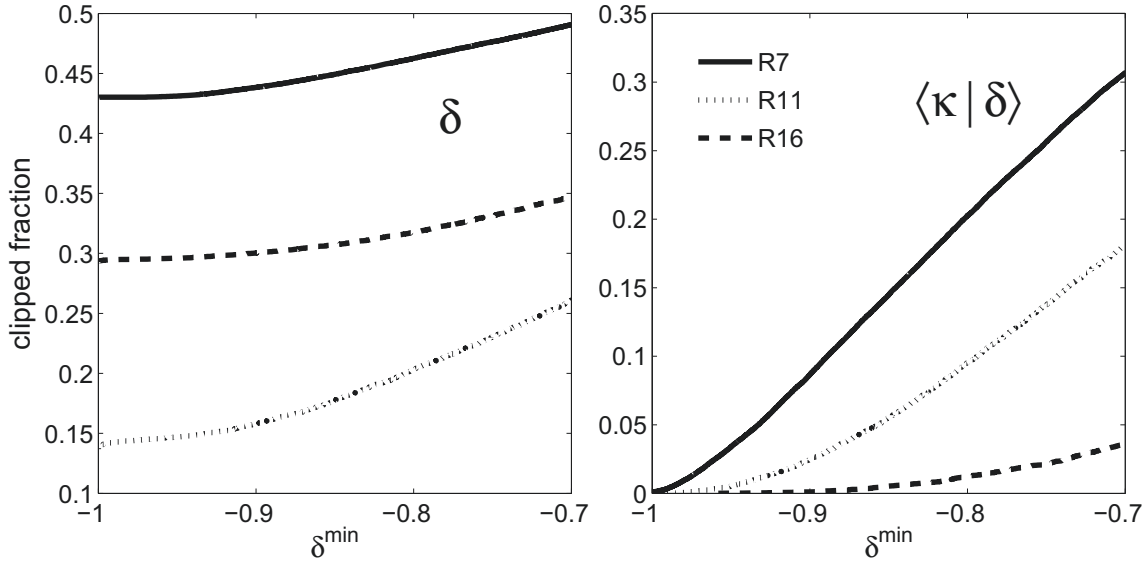

Figure 8.3: Fractions of cells with overdensities below δ^{min}. The left panel shows the distribution of the elements of $\boldsymbol{\delta}$ before cleansing. The right panel shows the same for the estimated signal $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$. The term δ^{min} is also used in the context of "clipped overdensities" where the $\delta_c(\boldsymbol{x}) = \delta(\boldsymbol{x})$ if $\delta(\boldsymbol{x}) > \delta^{min}$, and $\delta_c(\boldsymbol{x}) = \delta^{min}$ otherwise. This formulation is commonly used when calculating log-spectra $P_{\ln(1+\delta_c)}(k)$ to avoid taking the log of zero.

Many of the R16 cells at $z > 0.22$ separate into distinct "trails". Cells in each trail are geometrically related, either by lying along edges of the survey or being subjected to similar signal or noise features. For example, the trail in the upper right corner of the plot contains high redshift cells for which $\beta_{PS} < 1$. These extend beyond the boundary of the photometric footprint and have correspondingly lower $\langle n \rangle$.

The cleansing algorithm will shift the distribution of overdensities, but it should not significantly affect the total galaxy count. If it did, the overdensity of the entire survey volume would deviate from zero, in violation of the cosmological principle.

We find that the total count of DR6 MGS targets is largely invariant to our combination of data corrections. As shown in Figure 8.5, despite perturbations in individual redshift

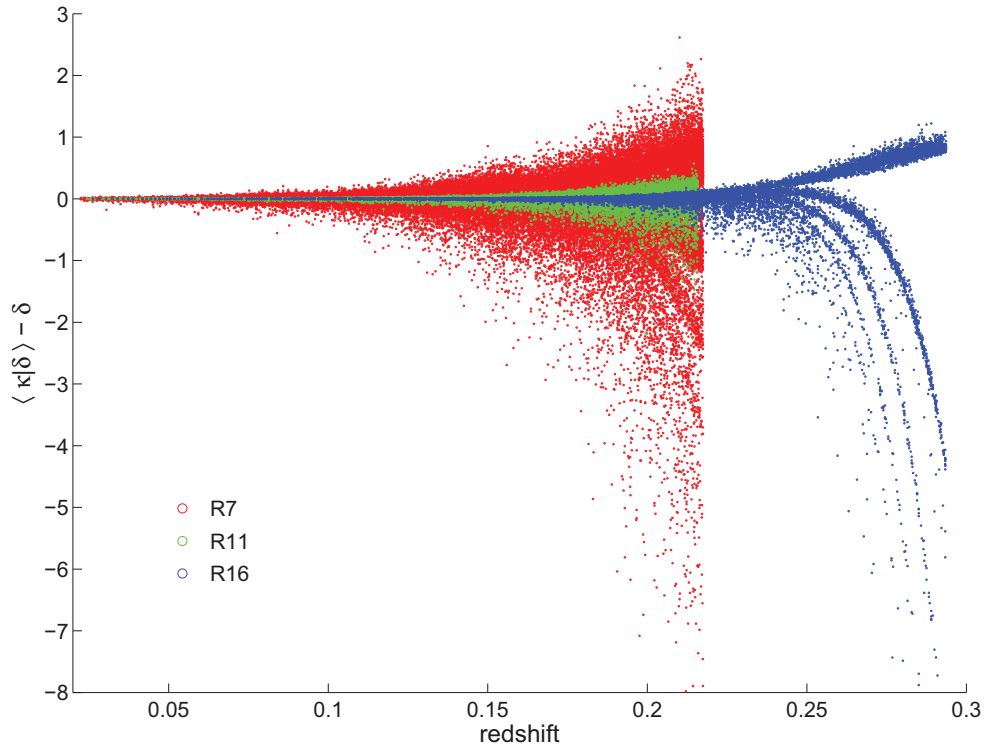

Figure 8.4: Scatter plot of changes in overdensity as a function of redshift. Each pixel represents a single cell from the R7 (*red*), R11 (*green*) or R16 (*blue*) set. The change in overdensity is defined to be the cleansed overdensity $\langle \kappa_i | \boldsymbol{\delta} \rangle$ minus the raw overdensity δ_i.

bins, the total number of galaxies within the survey is very nearly conserved. Rounded to the nearest integer the changes in number for R7, R11, and R16 are 507, 155, and -233. These represent percentage changes of 0.6%, 0.8%, and -1.4% respectively.

8.2 Power Spectra and Correlation Functions

With the census of overdensities complete, we turn our attention towards the cleansing's impact on the power spectra and 2PCFs. In this section, let $P_\delta(k)$ represent the power of

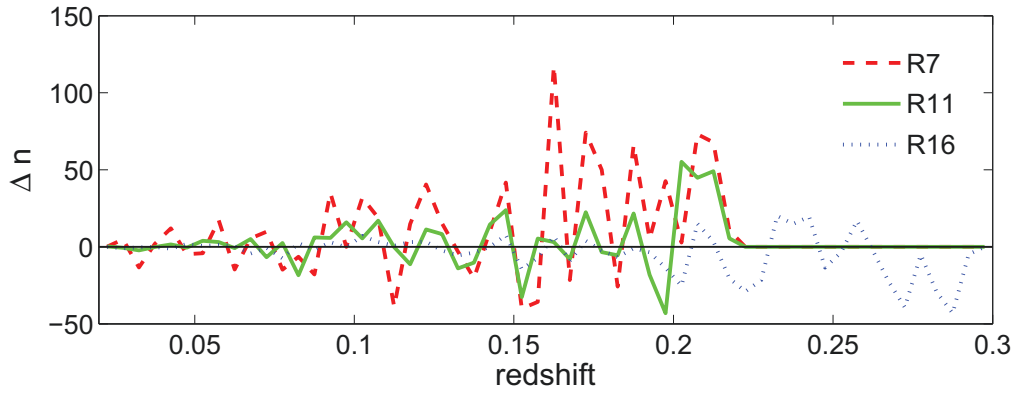

Figure 8.5: Change in the number of MGS galaxies after cleansing. Cells are organized into redshift bins of width $\Delta z = 0.005$. On the vertical axis $\Delta n(z) \equiv n_c(z) - n(z)$ where n is the number of galaxies within that redshift slice prior to cleansing and $n_c = \langle n \rangle (1 + \langle \kappa | \delta \rangle)$ is the number after cleansing.

the measured data prior to cleansing, and $P_{\langle \kappa|\delta \rangle}(k)$ represent the power of the estimated signal. $P_{\langle \eta|\delta \rangle}(k)$ and $P_{\langle \zeta|\delta \rangle}(k)$ will be used to denote the powers of the estimated zero-point and shot noises, respectively. We refer to $\langle \boldsymbol{\eta}|\boldsymbol{\delta} \rangle + \langle \boldsymbol{\zeta}|\boldsymbol{\delta} \rangle$ as the "estimated noise".

8.2.1 Recovered Power

We expect the power of the estimated noise to be relatively constant at intermediate k with a slight increase at the largest scales to account for zero-point effects. On small scales the power should fall off as a function of the spherical smoothing kernel. These features were first depicted in Figure 7.3 and can be seen in combination in Figure 8.6.

In Figure 8.7, the power of the raw overdensity data, the power of the estimated noise (represented by $P_{\langle \kappa|\delta \rangle}(k)$), and their respective fiducial models are presented. Overall, the cleansing algorithm performs well in recovering the power $P_\kappa(k)$ of the fiducial signal

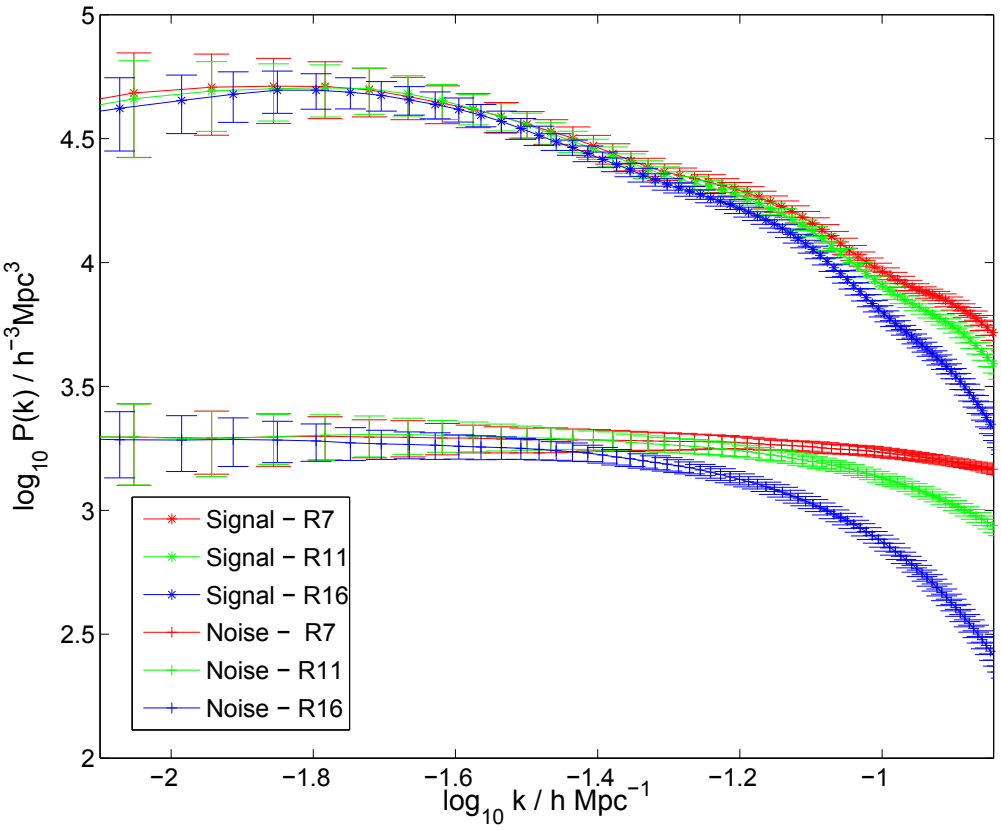

Figure 8.6: Comparison of power spectra of fiducial models as a function of cell size. $P_\kappa(k)$ are superimposed and differ in shape only due to the effect of the spherical window functions. The power of the combined shot plus zero-point noise (labeled *noise* in the legend) are also presented. To enhance overlap, $P_{\eta\zeta}(k)$ are scaled by factors of 0.9 and 0.04 for R11 and R16, respectively. Generation of these spectra were discussed in Chapter 4. Unscaled versions were originally presented in Figure 7.3.

for the R7 and R11 cases. The power of the estimated noise falls well within the range anticipated by the noise models. The structure of the estimated noise is that of a scale-invariant function convolved with a spherical window, much like the shot noise itself.

An uptick in the raw data's large-scale power for the R7 case is reflected in a corresponding uptick in the power of the estimated noise. This data uptick, which might result

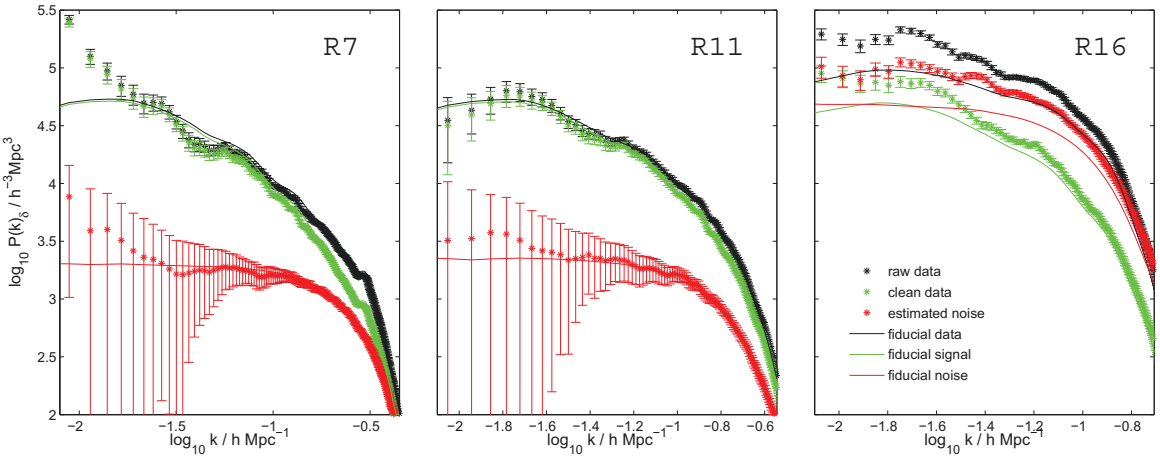

Figure 8.7: Power spectra of raw overdensity data $\boldsymbol{\delta}$ (before cleansing) and $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$ (after cleansing) are presented in black and green, respectively. The power of the estimated noise $\langle \boldsymbol{\eta} | \boldsymbol{\delta} \rangle + \langle \boldsymbol{\zeta} | \boldsymbol{\delta} \rangle$ is shown in red. For all three components, the average powers of the fiducial models are given by solid curves. The error bars represent 1σ variations in power generated from 250 overdensity realizations drawn from the fiducial models. Note that each set of error bars communicates the uncertainties within a fixed model, but not those *between* models.

from cosmic variance, is interpreted by the signal estimation algorithm as being a source of noise and is cleansed as such. A similar uptick in data power is missing in R11 and no corresponding uptick in the estimated noise power is observed.

The power measured across the R16 cells exhibits less agreement with the fiducial models than those constrained to $z < 0.22$. This could be the result of a number of different factors, but the most likely explanation is a discrepancy between the fiducial signal/noise models and the underlying fields from which the data are sampled. The signal model could weaken at larger redshifts or $n_{exp}(z)$ (from which the shot noise derives) could be inadequately constrained once the selection function drops below a certain limit, leading to an underestimation of shot noise power. This could also be indicative of a weakness in the counting method conclusions of Chapter 6. We must also concede the possibility of an un-

detected coding or computation error, though a focused search was unable to discover any. Regardless, $P_{\langle\kappa|\delta\rangle}(k)$ for the R16 cells still lies within 2σ of $P_{\kappa}(k)$ for most wavenumbers and is clearly preferable to $P_{\delta}(k)$ in terms of signal estimation.

We reiterate that the reported error bars represent the variances *within fixed* fiducial models, whereas the discrepancies in Figure 8.7 likely reflect a difference *between* models. We acknowledge that our error bars are smaller than those often reported in the literature. However, many of those studies (see e.g. Howlett et al., 2015; Eisenstein et al., 2005; Cole et al., 2005) largely rely upon the FKP method of error analysis. That method is largely restricted to linear power spectra under the assumption that galaxies are distributed in a Poissonian fashion. They also calculate power spectra using individual galaxy counts rather than overdensities in cells. These differences introduce incompatibilities into our respective error approaches. A unification of the two theories is probably possible, but a question we sidestep for now.

8.2.2 Effective Spectra

The cleansing process reduces power on all scales, resulting in a spectrum for $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$ that "passes the eye test" in terms of its resemblance to $P_{\kappa}(k)$. It is possible to be more rigorous in this determination, though, in order to strengthen the argument that the power removed was comprised of more noise than signal.

The approach relies upon the observation that the underlying galaxy clustering field is the same for all cell sizes. The only impact a cell's radius has on the measured power is the

effect of the spherical convolution. In principle, rescaling $P_\kappa(k)$ by the window function should restore it to the fiducial.

Both underlying noise fields, on the other hand, are dependent on cell size. All else being equal, smaller cells induce more shot noise. Their projections also intersect fewer PRIMARY SEGMENTs on average, reducing the offsetting effects of adjacent zero-points and enhancing systematic noise. While the noise fields are also convolved by cell sizes, scaling them by their cells' respective window functions will not restore them to a common fiducial spectra, since no common spectra exist. (Reference Figure 8.6 for a reminder of how these spectra compare.)

If the noise cleansing process is effective in generating a $P_{\langle\kappa|\delta\rangle}(k)$ that closely resembles $P_\kappa(k)$, then scaling $P_{\langle\kappa|\delta\rangle}(k)$ by the window functions should yield "fiducial" spectra that more or less overlap. Yet if $P_\delta(k)$ were scaled by the same window functions, we should expect less overlap due to the presence of noise.

One complicating factor is that the $P_{fid}(k)$ that seeded the signal models was a real-space matter spectrum. The simulated power spectra of Figure 7.3 were subjected to redshift-space distortions, separation distances calculated through the non-Euclidean Liske geometry, and subsequent convolution through the SDSS survey window function, all of which introduced k-dependent effects for which the spherical kernel alone cannot account.

To approximate the impact of these effects, we define an *effective kernel* $K_R(k)$ for each cell size R such that

$$P_\kappa(R,k) = K_R(k) P_{fid}(k), \tag{8.1}$$

where $P_\kappa(R,k)$ is the simulated signal power spectrum from Figure 7.3 and $P_{fid}(k)$ is the unconvolved fiducial MGS galaxy power spectrum in real-space. Figure 8.8 shows how the effective kernels and spherical window functions compare. The redshift-space distortions, Liske geometry, and SDSS survey windows serve to increase the effective kernel amplitudes by a factor of ~ 1.2, while largely preserving the shape.

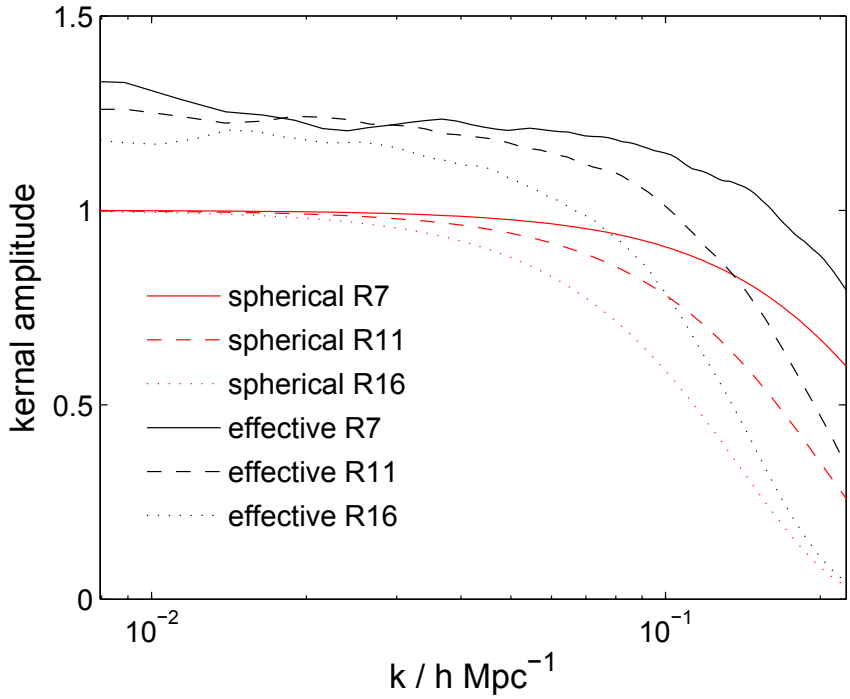

Figure 8.8: Comparison between the spherical window function $|W_R(k)|^2$ of equation (4.3) and the effective kernel $K_R(k)$. The effective kernel is calculated relative to the unconvolved fiducial *galaxy* power spectrum with a bias factor of $b = 1.2$.

To test the degree of overlap between the scaled spectra, we define the *effective fiducial spectra* $P_{ef}(R,k)$ for raw and cleansed data as

$$P_{ef}(R,k)_{\delta} \equiv P(R,k)_{\delta}/K_R(k),$$

$$P_{ef}(R,k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle} \equiv P(R,k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle}/K_R(k). \tag{8.2}$$

To determine the extent to which effective fiducial spectra overlap, the norm of the difference is taken for all three cell-pairing possibilities. We refer to this wavenumber-dependent quantity as the "degree of overlap." The results of these comparisons are illustrated in Figure 8.9. Curves with lower magnitudes reflect greater degrees of overlap.

In short, Figure 8.9 provides evidence that the cleansing process has produced an estimated signal $\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle$ that is closer to the true signal $\boldsymbol{\delta}_{\kappa}$ than the raw data $\boldsymbol{\delta}$ is — at least as conveyed through the power spectrum. This conclusion presents itself most forcefully in the comparison of cell sizes with the least (R11) and most (R16) noise. The degree of overlap between $P_{ef}(11,k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle}$ and $P_{ef}(16,k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle}$ is stronger for all k smaller than the characteristic cell size, than that between $P_{ef}(11,k)_{\delta}$ and $P_{ef}(16,k)_{\delta}$, and most often by a full order of magnitude or more.

A similar conclusion holds for the R7/R16 comparison, although here the degree of overlap for $P_{ef}(k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle}$ is even stronger on small scales. The degree of overlap weakens on the largest scales until reversing at the lowest measured values of k, although in this instance cosmic variance could be the culprit. The cleansing improvement between the R7 and R11 cases is relatively modest, but uniform for all but three measured values of k.

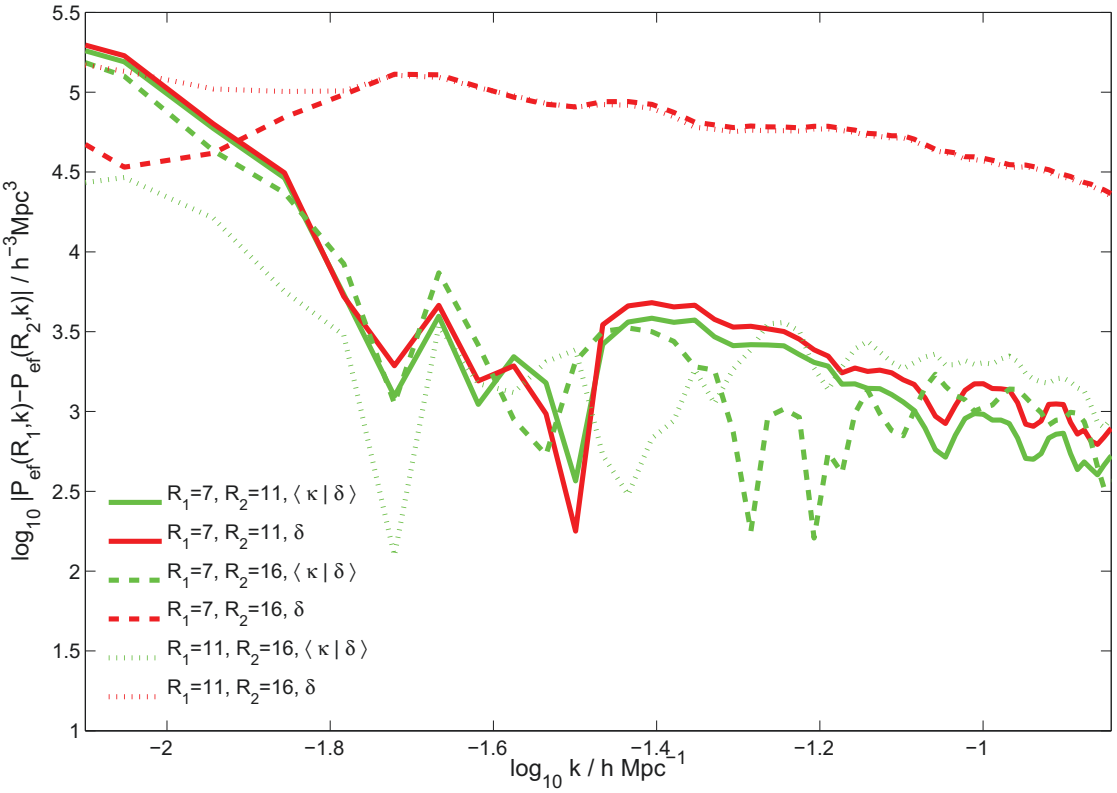

Figure 8.9: Degree of overlap for effective fiducial power spectra. Spectra overlap comparisons are conducted in pairs — R7/R11, R7/R16 and R11/R16 — with the absolute difference of P_{ef} presented on the vertical axis. Degrees of overlap amongst $P_{ef}(R,k)_{\langle\boldsymbol{\kappa}|\boldsymbol{\delta}\rangle}$ are plotted in green and represent the extent to which $P_{\langle\kappa|\delta\rangle}(k)$ overlap after "deconvolving" with the effective kernels. Values of zero on the vertical axis indicate perfect overlap. Degrees of overlap among raw data power $P_{\delta}(k)$ after "deconvolving" are shown in red. Put simply, curves of lower magnitude indicate better overlap.

The "overlap improvement" going from $P_{\delta}(k) \rightarrow P_{\langle\kappa|\delta\rangle}(k)$ is greater between R16/R11 and R16/R7 than between R7/R11. This offers another metric by which to conclude that the positive impact of noise reduction is most pronounced in the R16 case. This coincides with expectations given that this high redshift data set contains significantly more noise (see Table 4.6). Not only has signal estimation been successful at recovering the true clustering

power, but the improvement is greater when more noise is present.

These results are encouraging since there are several factors that inhibit an exact overlap. First, given that the derivation of $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$ relies upon three best-guess fiducial models, it is almost certainly imperfect. Moreover, the degree of overlap statistic depends upon the signal clustering model through the effective kernel. If the true underlying signal field differs from this model, then $K_R(k)$ cannot be expected to precisely map the signal back to the fiducial spectrum. This difference might also be impacted by the bias factor b which, in addition to its $\sim 10\%$ uncertainty, also fails to take and scale- or luminosity-dependent biases into account.

8.2.3 Adjusted 2PCF

To assess how cleansing affects the excess probability of finding two galaxies separated by a distance r, we examine the simple one-dimensional two-point correlation function. To generate $\xi(r)$ we integrate $P(k)_\delta$ and $P(k)_{\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle}$ over k-space through equation (3.27). The results are presented together in Figure 8.10. The differences are shown separately through the first $100\,h^{-1}\mathrm{Mpc}$ in Figure 8.11 and together through the first $30\,h^{-1}\mathrm{Mpc}$ in Figure 8.12.

The primary impact of cleansing the R7 and R11 cells is a reduction of $\xi(r)$ on scales $r \leq 30\,h^{-1}\mathrm{Mpc}$. Beyond this scale, the change approaches zero. There are separations r at which cleansing does restore $\xi(r)$, but these changes are 2+ orders of magnitude smaller than the original changes. The locations of positive adjustments to the 2PCF differ as a

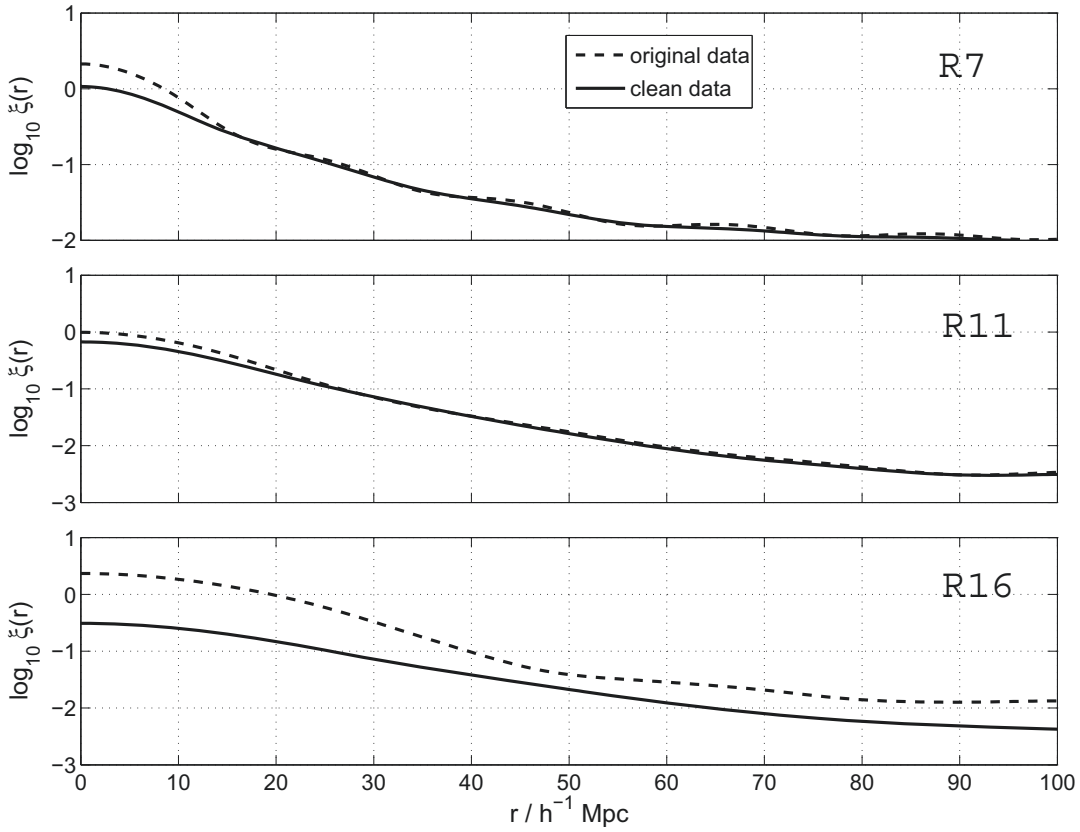

Figure 8.10: Two-point correlation functions of spherically convolved MGS data before and after cleansing. Correlation functions are calculated through Fourier transforms of the power spectra of the data. From top to bottom, these are the 2PCFs of cells of radii 7, 11 and 16 $h^{-1}\mathrm{Mpc}$.

function of cell size.

The rank order of the magnitudes of the reductions in $\xi(r)$ at small separations is the same as the rank order of noise magnitudes in cells. As with the R7/R11 cases, the reduction in the 2PCF for R16 cells at on scales $r \lesssim 30\,h^{-1}\mathrm{Mpc}$ is greater than it is elsewhere. Unlike those cases, however, this reduction continues through the distances at which galaxy positions are no longer correlated.

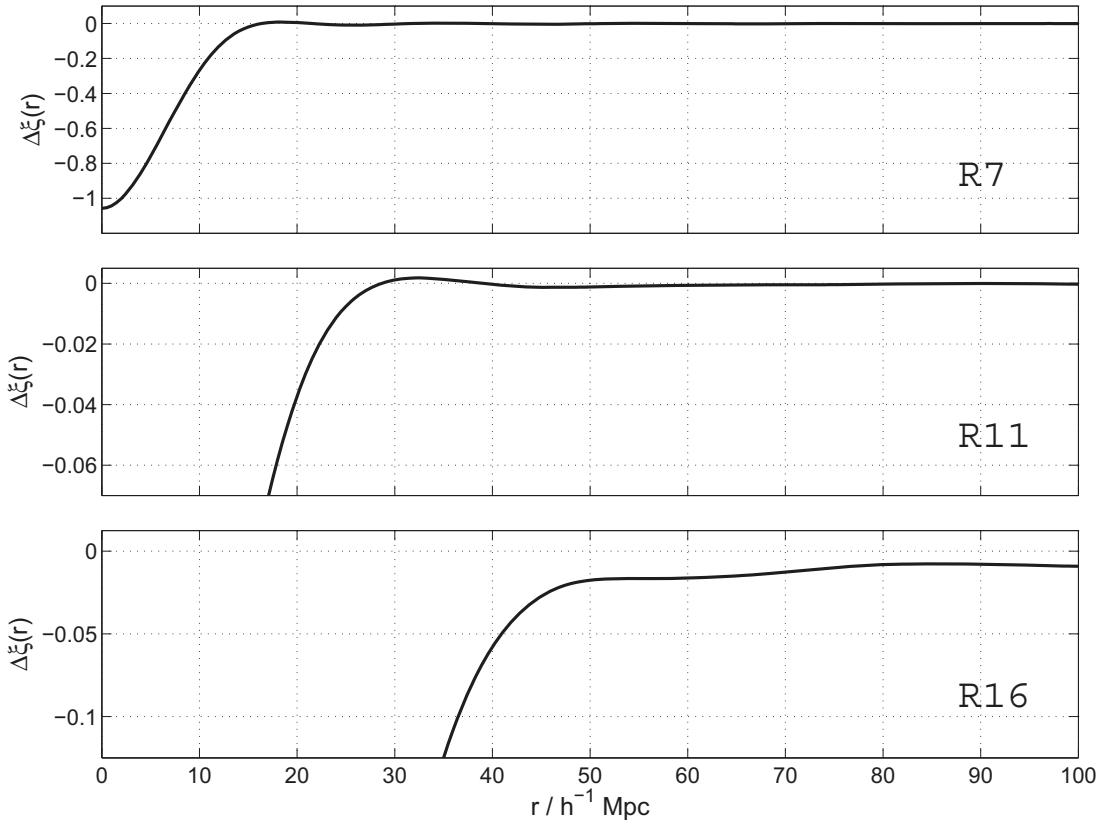

Figure 8.11: Differences between the 2PCFs of the clean data and the original data as presented in Figure 8.10.

8.2.4 Estimated Shot and Zero-Point Noise

We conclude this chapter by attempting to estimate the shot noise and systematic zero-point noise using equations (7.17) and (7.19). The aggregate noise was well estimated for the R7 and R11 cells, while the shape of the noise spectrum was recovered for the R16 cells (see Figure 8.7). However, as Figure 8.13 illustrates, when handling the real data our algorithms are far less effective at breaking the degeneracy between the two sources of noise.

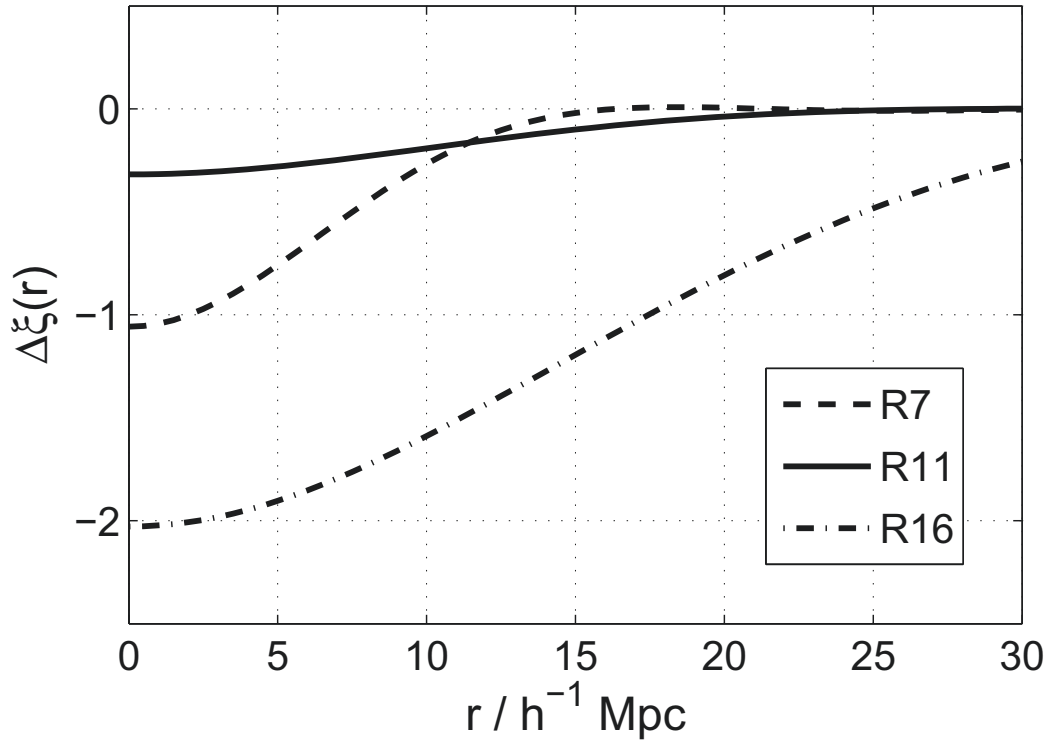

Figure 8.12: Alternative view of Figure 8.11 focused on the first $30\,h^{-1}\mathrm{Mpc}$.

According to our fiducial models, shot noise is approximately 100 times more powerful than zero-point noise on small and intermediate length scales, while the zero-point noise becomes relatively more powerful with decreasing k. The powers of the estimated noise components, on the other hand, differ in two significant ways — $P_{\langle\zeta|\delta\rangle}/P_{\langle\eta|\delta\rangle}$ is largely scale invariant and $\langle\boldsymbol{\eta}|\boldsymbol{\delta}\rangle$ contains slightly *more* power than $\langle\boldsymbol{\zeta}|\boldsymbol{\delta}\rangle$.

This result is illuminating. As reported during the simulated diagnostic tests in §7.6.2, we not only successfully estimated $\langle\boldsymbol{\eta}|\boldsymbol{\delta}\rangle$ across a range of realizations, but for the longer PRIMARY SEGMENTs we were able to make useful predictions about the magnitudes of photometric offsets. In that scenario, the underlying signal and noise fields were known

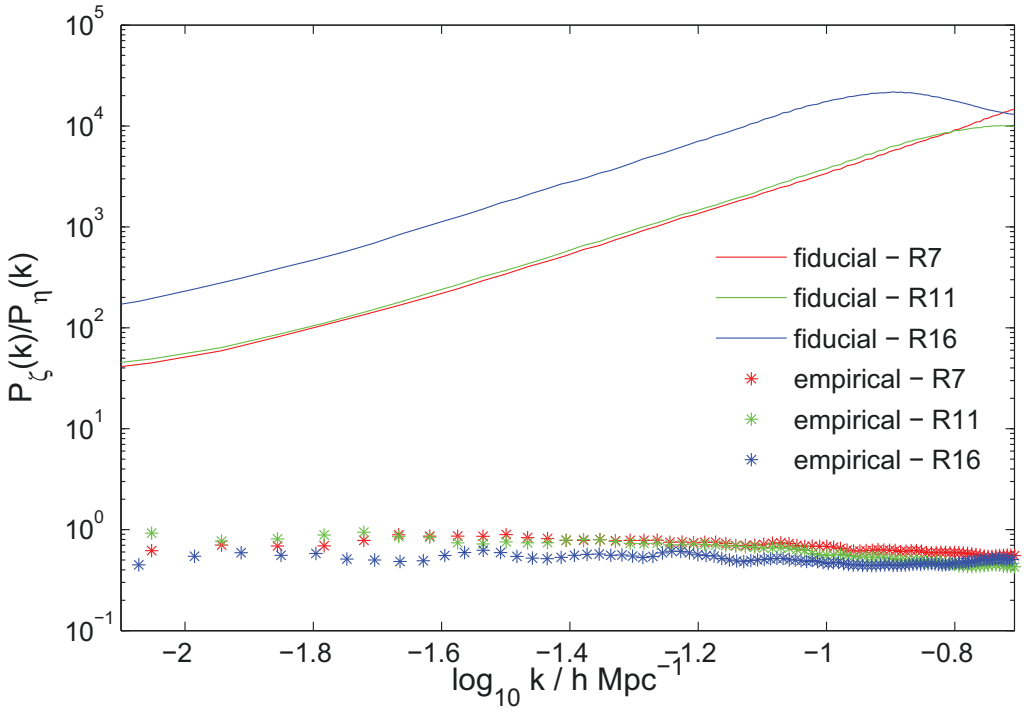

Figure 8.13: Power of shot noise relative to that of zero-point systematic noise. The solid curves display the ratio of $P_\zeta(k)$ to $P_\eta(k)$ as quantified through the fiducial models of Figure 7.3. The starred points reveal the ratio of $P_{\langle\zeta|\delta\rangle}(k)$ to $P_{\langle\eta|\delta\rangle}(k)$, i.e. the relative power of the estimated noise components. Error bars are omitted for clarity.

perfectly, yet the estimates still pushed the limits of our computational abilities.

In this empirical case, conditions were far less idealized. Our fiducial signal model almost certainly differs in some appreciable way from the true clustering field. Discrepancies in the R16 power spectra of Figure 8.7 further suggest modeling errors for one or both sources of noise. Such incongruities between models and reality not only strain efforts to estimate shot noise, but make it virtually impossible to detect effects that are 10 to 100 times smaller. Through this, we begin to see the limits of these noise estimation algorithms.

As with all Bayesian analyses, the quality of the posterior distributions are only as good

as the priors that seed them. When the priors were known perfectly, the techniques introduced in this dissertation proved themselves capable of successfully estimating sub-percent level systematic effects. When handling real data, these techniques were capable of reliably recovering $P_{\langle\kappa|\delta\rangle}(k)$. This was true even in the R16 case where $P_\delta(k)$ exceeded $P_\Gamma(k)$ by a significant margin. It is encouraging that the aggregate noise could be quantified, even if the size of each individual component could not.

A logical follow-up analysis would repeat these tests using a range of different cosmological signal and fiducial noise models. It might be possible to iteratively perturb relevant model parameters until the estimated signal and noise components match the statistical properties of the input models. This could simultaneously provide signal and noise estimates while constraining the models' parameter spaces. Of course, we must bear in mind that throughout this entire chapter, each cell set has only considered a single $\boldsymbol{\delta}$-vector. Under such circumstances cosmic variance will always play a part.

Chapter 9

Conclusion

Throughout the course of this dissertation, we introduced solutions to three types of problems that inhibit precision cosmology: errors in the survey footprint, accounting for galaxies without redshifts, and statistical/systematic noise. Here we summarize the results of each of these investigations and issue some closing remarks.

9.1 Footprint Corrections

Chapter 2 expanded and clarified numerous issues related to geometric definitions of the SDSS photometric and spectroscopic footprints. The concept of PRIMARY SEGMENTs was introduced and later utilized to model zero-points. SECTORs were reviewed en route to improving the survey's spectroscopic completeness.

All problems with the SDSS footprint were found by eye through visualizations. To en-

sure the errors were legitimate and not, for example, the result of cosmological variance, we compared DR6 MGS targets' positions against DR7 footprint definitions. We discovered that the DR7 footprint contained "hidden information" about DR6 that was indispensable to improving the survey boundary.

First, we found that five regions within the DR6 photometric footprint contained zero targets. Another STRIPE in the southern hemisphere was contaminated by ambiguous, overlapping region definitions (see Table 5.1). Removing these regions from the union of PRIMARY SEGMENTs reduced its area by 19.7 deg^2 , or 0.24% of the total. Despite their small aggregate size, we learned that these trouble regions could induce an error in $\langle n \rangle$ that exceeded 60% for some R7 cells.

Second, we uncovered a number of regions within the spectroscopic footprint that were erroneously included or excluded. Five regions contained MGS galaxies but had been excluded from the union of SECTORs (see Table 5.2). These regions were restored to the footprint. Another set of areas were included but should have been (and subsequently were) removed (see Figures 5.14, 5.18 and 5.19).

Third, we discovered over 300 SECTORs near the spectroscopic survey boundary that were grossly spectroscopically undersampled relative to the rest of the survey. As summarized in Table 5.3, only 25% of MGS targets within the SECTORs we trimmed away had their spectra taken. This compares to 79.5% elsewhere.

The sum of these changes has resulted in the most accurate description of the SDSS DR6 survey footprint that has ever been created. A 3.6% reduction in the spectroscopic

footprint area led to a 2.2% increase in the spectroscopic completeness of the entire survey. These changes significantly increased the accuracy of overdensity measurements within cells on the survey's boundary, without removing an inordinate amount of useful spectroscopic data in the process.

9.2 Noise Cleansing

We also introduced a general method to reduce the impact of systematic errors in large data sets. We showed that when both the signal and noise are generated by Gaussian processes, a simple analytic solution to the optimal Bayesian signal estimate exists. We reported that such a solution is computable in a reasonable amount of time if one's space is discretized into tens of thousands of dimensions or fewer. We verified these conclusions using an empirical MCMC process, which we also offer as an alternative solver in instances where no analytic solution exists.

In summary, our basic solution framework is:

1. Discretize the problem into ∼60,000 dimensions.

2. Derive covariance matrices for the signal you wish to recover and combined sources of noise within those dimensions.

3. Solve the W-space eigenproblem.

4. Using equation (7.11), solve for $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$.

We tested our method by modeling photometric calibration zero-point offsets within the Sloan Digital Sky Survey. Using MGS galaxies, we calibrated the Schechter luminosity function and derived a relation $f(z)$ that gave the fractional change in galaxies expected in the absence of clustering as a function of the sample's limiting magnitude. By specifying the geometric intersections between DR6 PRIMARY SEGMENTs and our cells, we were able to map zero-points to galaxy overdensities.

Using simulated galaxy overdensities from the MGS, we demonstrated that our signal estimation method is capable of reducing the noise variance in cells due to shot noise and systematic noise. We found that for cells of radii $7h^{-1}\mathrm{Mpc}$ and $11h^{-1}\mathrm{Mpc}$ tightly packed in the redshift range $0.02 < z < 0.22$ the decreases in noise variances were, respectively, 47.9% and 29.5%. This result is consistent with the intuitive idea that discretizing space over a larger number of higher resolution dimensions is preferable to doing so over a smaller number of dimensions. This is most likely due to the fact that more dimensions permit a greater number of correlations between cells, enhancing the ability to pick out the noise components.

For spheres of radius $16h^{-1}\mathrm{Mpc}$ in the redshift range $0.02 < z < 0.30$, the cleansing was stronger. Even though the overdensity measurements in those cells were more contaminated by noise than in the R7 and R11 cases, our framework performed well, reducing the noise variance by 82.3%. This result came even though the number of R16 dimensions was fewer (15,166 cells versus 78,845 and 19,737 for R7 and R11 respectively).

We also demonstrated that our cleansing framework improved individual overdensity

measurements in a majority of cells, regardless of radius. For R7, R11, and R16 respectively, the percentages of cells that saw improvements were $60.4 \pm 0.2\%$, $57.1 \pm 0.4\%$ and $65.1 \pm 0.4\%$. Moreover, using the vector 2-norm as a metric we found that our estimated signal $\langle \boldsymbol{\kappa} | \boldsymbol{\delta} \rangle$ did a better job of approximating the true signal $\boldsymbol{\delta}_\kappa$ than the raw data did.

We presented a new estimator $\hat{P}_{\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle}$ for the power spectrum and compared it against the spectra of the raw data $P_{\boldsymbol{\Gamma}^{(\tau)}}$ and true signal $P_{\boldsymbol{\kappa}^{(\tau)}}$. As shown in Figure 7.4, $\langle \boldsymbol{\kappa} | \boldsymbol{\Gamma}^{(\tau)} \rangle$ did a better job of recovering the signal power than did $\boldsymbol{\Gamma}^{(\tau)}$. The improvement was most significant for the R16 cells, which contained the greatest amount of noise.

Methods to predict underlying levels of shot noise and photometric zero-point noise were presented. Being of higher amplitude, shot noise vectors proved easier to estimate than zero-point overdensities. In both cases, our method did well to predict both the sign of the error and its relative size between cells.

Overdensity errors introduced by zero-point offsets were more effectively estimated when cells intersected long PRIMARY SEGMENTs. Because our method utilizes statistical correlations between cells, a greater number of cell/PRIMARY SEGMENT intersections provide more constraints and consequently enable better estimates.

We found that utilizing either one of our noise estimates was preferable to adopting the default assumption that the noise in each cell takes its most likely value of zero. Our noise estimators are "safe" in the sense that when the errors are small and/or less well correlated with other cells, they return conservative values for the underlying noise.

We reemphasize that the methods of Chapter 7 operated on a random, systematic noise

component that contributed less than one ten-thousandth of the variance of the total data vector, and yet useful predictions were still possible. Of course, there will be a great many problems for which systematics are less well-constrained than they were for the SDSS zero-points. It is worth recognizing that if σ_m tripled, the noise variance would have increased by an order of magnitude. And as shown through the efficacy of the shot noise estimator, our method grows considerably more powerful as S/N drops.

9.3 Counting and Cleaning

We tested our cleansing algorithm on real MGS data gathered in the fashion described in §2.3. Approximately 20% of MGS targets within the improved spectroscopic footprint lack spectra. Unable to directly quantify these objects' radial depths through their redshifts, we spent Chapter 6 researching alternatives that could recover the true overdensities in cells.

We divided the survey footprint into three types of regions. *Interspersed regions* lay inside the improved spectroscopic footprint where the average angular distance between MGS galaxies and objects was small. *Dark regions* were those that existed outside the improved spectroscopic footprint but within cells' circular projections. Almost all targets within these regions lacked spectroscopic redshifts, though their proximity to the spectroscopic footprint made their average angular separation from MGS galaxies more manageable. *External regions* were those that did not lie in close proximity to the spectroscopic footprint. None of the objects within them possessed measured spectra.

We tested nine counting techniques to assess which were most effective at recovering cells' true galaxy count, overdensity and overdensity squared. These techniques were divided into three classes: discrete counting methods (ignore, nearest neighbor, SED photo-z, D1 photo-z), probabilistic smearing methods (selection function, 2PCF, SED photo-z, D1 photo-z) and scaling.

Bootstrapping was used to simulate mock MGS catalogs separated by region type. The counting methods were applied and compared against reality. For all three region types, ignoring objects in high-redshift cells was the most preferable option, revealing that none of the methods tested were particularly adept at handling volumes with low galaxy densities. The redshift at which ignoring objects became preferable increased with cell size.

With few exceptions, scaling galaxy counts in interspersed regions by each cell's spectroscopic completeness fraction proved to be the best approach at low redshifts. A similar result held for R7 cells in dark regions, though the transition to favoring ignoring galaxies occurred at a smaller redshift. For larger cells at low redshifts in dark regions, a combination of photometric redshift smearing and scaling dominated, as reported in Table 6.2.

No method emerged as being significantly better than the others when attempting to count objects in external regions (see Tables 6.3 and 6.4). Although the process of smearing photo-z's was able to recover the radial distribution of galaxies more effectively than the selection function, we could find no way to utilize these regions for precision cosmology.

The optimal counting strategies for each region type, redshift and cell size were applied to the MGS data. In Chapter 8 we subjected that data to our cleansing method. We found

that cleansing causes the overdensity histogram to contract towards zero, reflecting the principle that a cell with a negative (positive) overdensity is more likely to be comprised of negative (positive) noise.

The power spectra of the cleansed data more closely matched that of the fiducial signal spectra than the fiducial data spectra. This suggested that cleansing had removed noise. Using an effective kernel function, we reinforced this conclusion by quantifying the degree of overlap between "deconvolved" power spectra. We discovered that the power spectra of the estimated signal had been scrubbed of much of the noise drawn from cell-size-dependent underlying noise fields.

As cited in the text, there are no fewer than 19 separate photometric redshift codes currently in use. Some assume the probability distributions of photo-z's are Gaussian (as we have), while others generate distributions that are more individualized to the galaxy in question. Our conclusions are reached using one of the simplest models for photo-z's, one that we expect can be outperformed by other codes. Therefore we encourage the reader to view these conclusions as something of a lower limit of what probabilistic smearing can offer. These results should serve as an initial snapshot of how various basic counting methods perform when measuring overdensities in cells.

9.4 Closing Remarks

We would like to reiterate that the Bayesian noise cleansing framework developed and exercised here is fairly general. It should work for a wide range of problems involving systematic errors in large data sets, both inside and outside of astronomy. Provided the signal and noise covariances can be specified, our framework should lessen the need for time-consuming, expensive, or potentially impossible workarounds, e.g. repeated observations of a system.

Even though systematic effects did not dominate counting statistics in this particular problem, as survey volumes and galaxy counts increases, shot noise will drop, leaving systematic errors as the largest source of uncertainty in one's measurements. As computation power improves, so too will the number of discretized dimensions one can separate space into, leading to further performance gains. In other words, this method will grow more powerful with time.

We recommend using tens of thousands of dimensions to discretize over if you have on the order of 40 processors working in parallel and 100+ GB of contiguous memory for the required matrix inversions, and diagonalizations. (The Metropolis-Hastings step is similarly expensive, but the parallelization and contiguous memory requirements are lessened.) Too few discrete elements will weaken one's results while too many will make the computational costs too expensive.

A couple years before the inception of this project, our research group operated computers with 4 processors and 24 GB of contiguous memory. By its completion, we were

working on multiple machines each with 32 processors and 512 GB of contiguous memory. The additional memory permitted the maximum matrix dimension to increase by a factor of approximately 4.6.

Advancements like these were indispensable for performing the types of analyses discussed in this dissertation. Matrix operations such as inversions and diagonalizations are computationally expensive and must be parallelized. Even with our current computers, calculating the $\mathbf{W}$ matrices for the R7 cells took almost a week. Were we to utilize the full 512 GB, we could increase the number of cells to somewhere between 150,000 and 200,000 at which point the limiting factor becomes how quickly those matrix operations can occur. On present architecture this would require approximately two months.

While we have not explicitly addressed it in these pages, we recognize that the quality of the solutions depends on the accuracy of the signal and noise covariance matrices utilized. It is easy to imagine quantifying the uncertainty of one's solutions using an iterative process. Raw or partially cleansed data can be used to determine the probability space within which combinations of cosmological parameters reside. Repeating the analysis with various combinations (i.e. various fiducial models) can quantify how sensitive one's estimates are to the perturbations in the model.

If one were studying power spectrum features on translinear scales or smaller, for example, a change in the fiducial signal model might only require adjusting a limited number of signal covariance matrix elements corresponding to cells in close proximity. In Appendix H, we work through a technique for efficiently inverting matrices after perturbations. Ad-

ditional research may reveal efficient algorithms to streamline such a process.

We do not know precisely in what positions investigators will find themselves in the future. Research capacity is as much a function of funding availability and collaboration politics as it is about the state of computational architecture. We believe we have only scratched the surface of what our new scientific paradigm — eScience — has to offer. And as eScience matures, methods like those described in these pages will be critical tools to further establish cosmology as a high precision science.

Appendix A

Appendix A – Distance Measures

A.1 Distances to Objects

The Universe is constantly expanding, pulling galaxies ever further away from each other. The size of the younger Universe relative to today is given through the *scale factor* $a(t)$. Using the subscript 0 to denote "the present point in time" we set $a_0 = 1$.

For example, consider two objects that are today separated by a distance $\chi = 1000\,\mathrm{Mpc}$ and whose distance from one another depends solely on the expansion of the Universe. In the past when, say, $a(t) = 0.6$, the physical separation between the two would have been just $r(t) = 600\,\mathrm{Mpc}$ where

$$r(t) = a(t)\chi. \tag{A.1}$$

The term $r(t)$ in (A.1) is referred to as the "proper distance" and is equal to the true physical

separation between objects. χ is known as the "comoving distance" and is fixed in time to help calibrate the distance scale. Note that today proper distance and comoving distance are equal. For all previous points in time $r(t) < \chi$.

Because comoving distance is defined by the proper distance *today* and a_0 must always equal 1, the implication is that astronomers must continuously recalculate χ for each object. In principle, that concern has merit. In practice, though, the relative change in proper distance goes as $\Delta r/r \cong 2.25 \times 10^{-18}\, t$ where t is in seconds. This means that over the course of a human lifetime, the proper distance (and by extension the comoving distance as well) only changes by about one part in a billion. Given that astronomers are nowhere close being able to achieve that kind of measurement accuracy, this concern can more or less be safely ignored.

Differentiating equation (A.1) at time t_0 while noting that χ is constant in time yields *Hubble's Law*, $v = H_0 r$. This relates the proper recessional velocity v to the proper distance through the *Hubble parameter*, $H(t) = \dot{a}(t)/a(t)$. We say that a galaxy whose Hubble velocity equals the speed of light is at a distance called the *Hubble distance*,

$$d_H(t_0) \equiv c/H_0. \tag{A.2}$$

For the most part, the comoving distance χ between two distant objects does not change much with time. But for those that do, velocities measured with respect to the comoving frame are called *peculiar velocities*. Most large mass concentrations have small peculiar velocities compared to their Hubble velocities. Its presence is the motivation behind using

a redshift-space correlation function as described in §3.4.

As a result of the Doppler effect, the true wavelength of light λ_0 emitted from a galaxy receding from Earth will be shifted to a longer observed wavelength λ_{obs}. The ratio of these quantities defines the *redshift* z of the galaxy, adjusted such that our galaxy lies at $z = 0$,

$$\lambda_{obs} = \lambda_0(1 + z). \tag{A.3}$$

This also serves as an alternate expression for the scale factor,

$$a(t) = \frac{1}{1 + z}. \tag{A.4}$$

Through integration of the Robertson-Walker metric (an equation which describes the length of a line element in an expanding Universe in flat space), it can be computed that

$$\chi(t_e) = c \int_{t_e}^{t} \frac{dt'}{a(t')}, \tag{A.5}$$

where t_e is the time of emission of the photons and t is the time now. For practical reasons it is preferable to express equation (A.5) in terms of observables. Rearranging equation (A.4) gives

$$z = \frac{a_0}{a(t)} - 1. \tag{A.6}$$

Differentiating equation (A.6) with respect to t yields $dt/a(t) = -dz/(a_0 H(z))$. The

quantity $H(z)$ is the Hubble parameter at the time an object at redshift z emitted the light. This allows us to rewrite equation (A.5) for the comoving distance as

$$\chi(z) = -\frac{c}{a_0}\int_z^0 \frac{dz'}{H(z')} = d_H \int_0^z \frac{dz'}{H(z')}. \tag{A.7}$$

The time dependence of the Hubble parameter is derived through the Friedman equation (an equation that can be thought of as a conservation of energy statement in an expanding Universe) such that

$$H(z) = H_0(1+z)\sqrt{1+\Omega_m z + \Omega_\Lambda\left(\frac{1}{(1+z)^2}-1\right)}. \tag{A.8}$$

Peebles (1993, pp 310–321) introduces the function

$$E(z) \equiv \sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}, \tag{A.9}$$

where Ω_m, Ω_k, and Ω_Λ are the density parameters of matter, curvature and dark energy respectively.[1] This function plays a role identical to that of $H(z)$.

Using equations (A.7) and (A.9) to calculate comoving distances as a function of redshift requires the selection of a particular cosmological model. For all work located within these pages we employ equation (A.9) using a standard flat cosmology where $\Omega_m = 0.3$, $\Omega_k = 0$, and $\Omega_\Lambda = 0.7$. In evaluating d_H we assume $H_0 = 100\,h(\mathrm{km/s})/\mathrm{Mpc}$ such that

[1]The density parameters equal the ratio of the energy density of each component to the critical energy density of the Universe. The latter term is defined to be the density required to make the Universe spatially flat.

our comoving distances $\chi(z)$ are given in terms of $h^{-1}\mathrm{Mpc}$. The variable h is referred to as the *dimensionless Hubble parameter* and is assumed to equal $h = 0.7$ throughout.

There are two other distance measures referenced within these pages. The first is the *angular diameter distance* $d_A(z)$. This is calculated by combining the physical size r of an object with its apparent angular radius θ (in radians),

$$\theta = \frac{r}{d_A} = \frac{r(1+z)}{\chi}. \tag{A.10}$$

In a flat Universe, total comoving volume out to redshift z is $V_\chi(z) = (4\pi/3)\chi(z)^3$. The comoving volume element in solid angle $d\Omega$ is

$$dV_\chi(z) = d_H \frac{(1+z)^2\, d_A^2}{E(z)} d\Omega\, dz = d_H \frac{\chi(z)^2}{E(z)} d\Omega\, dz. \tag{A.11}$$

An object with luminosity L will be observed on earth as having a flux $f = L/(4\pi d_L^2)$ where d_L is the *luminosity distance*. The luminosity distance is related to the comoving distance in the following way,

$$d_L(z) = \chi(z)(1+z). \tag{A.12}$$

Because the human eye has a logarithmic response to light, astronomers historically have used a logarithmic flux measure, *magnitude*, frequently denoted with variables m and M. The *absolute magnitude* M of an object is calibrated to its luminosity through fixed standards M_0 and L_0 such that

$$M - M_0 = -2.5 \log(L/L_0)\,, \tag{A.13}$$

$$L/L_0 = 10^{(M_0 - M)/2.5}, \tag{A.14}$$

and

$$\frac{dl}{dM} = \left(-\frac{\ln(10)}{2.5}\right) 10^{(M_0 - M)/2.5}, \tag{A.15}$$

when $l \equiv L/L_0$.

The *apparent magnitude* m of an object is defined similarly, but as function of flux f rather than intrinsic luminosity L,

$$m - m_0 = -2.5 \log(f/f_0)\,. \tag{A.16}$$

For a single object as seen from two different distances, $L = L_0$ and

$$m - m_0 = -2.5 \log\left(d_0^2/d^2\right) = -5 \log(d_0/d)\,, \tag{A.17}$$

where d is by definition the luminosity distance. The absolute magnitude M is defined to equal the apparent magnitude when $d = 10\,\mathrm{pc}$,

$$m - M = -5 \log\left(\frac{10\,\mathrm{pc}}{d}\right) = -5 + 5 \log(d). \tag{A.18}$$

The quantity $m - M$ is referred to as the *distance modulus*. If distance is measured in terms of Mpc, as is more often the case,

$$m - M = -5 \log\left(\frac{10^{-5}\,\mathrm{Mpc}}{d}\right) = 5 \log(d) + 25. \tag{A.19}$$

Note that the SDSS uses multiple measures of flux (e.g. Petrosian, PSF, fiber) and therefore reports multiple magnitudes for each object. Unless otherwise noted, this analysis uses the r-band Petrosian magnitude r_P.

A.2 Distances Between Objects

The correlation function $\xi(r)$ depends on the distance between galaxies, but the magnitude of r cannot be measured directly. It must be calculated using the galaxies' redshifts z_1 and z_2 and their angular separation α. Converting redshifts to distances then solving for r using a simple Euclidean approach is insufficient, however.

Suppose Galaxy 2 emits a photon towards Galaxy 1. At the instant Galaxy 1 receives the photon, it emits one of its own bound for earth that is later measured at z_1. For the purposes of the correlation function, the relevant causally connected distance is $r = \chi_2'$, the comoving separation between Galaxies 1 and 2 at the moment Galaxy 1 receives and emits its photons.

The solution for χ_2' (and z_2', the redshift of the photon received by Galaxy 1) was worked out by Liske (2000, pp. 557–561) whose results we summarize here. These derivations as-

sume a homogeneous Friedmann (zero-pressure) cosmology with no cosmological constant ($\Lambda = 0$). Of course this is not strictly true, but since the MGS occupies a relatively small volume of redshift space ($z < 0.30$) and correlations between distant volumes are quite small anyhow, this approximation should be acceptable.

The solution below incorporates our continued assumption of a flat Universe. The values in equations (A.20) through (A.27) are needed for the final distance calculation.

$$P_+ = \frac{1}{q_0}\left[(q_0-1)\left(\sqrt{1+2q_0z_1}+\sqrt{1+2q_0z_2}-1\right)+\sqrt{(1+2q_0z_1)(1+2q_0z_2)}-q_0\right], \tag{A.20}$$

$$P_- = \sqrt{1+2q_0z_2}-\sqrt{1+2q_0z_1}, \tag{A.21}$$

$$P^2 = \frac{1+z_1}{4(1+z_2)}\frac{1}{1+2q_0z_1}\left(P_+^2\sin^2\frac{\alpha}{2}+P_-^2\cos^2\frac{\alpha}{2}\right), \tag{A.22}$$

$$q_1 = q_0\frac{1+z_1}{1+2q_0z_1}, \tag{A.23}$$

$$z_2' = \frac{2P^2}{(q_1-2P^2)^2}\left(1+q_1-2P^2+\sqrt{\frac{q_1^2}{P^2}+1-2q_1}\right), \tag{A.24}$$

$$H_1 = H_0(1+z_1)\sqrt{1+2q_0z_1}, \tag{A.25}$$

$$a_1 = \frac{a_0}{1+z_1}, \tag{A.26}$$

$$\chi_2' = \frac{c}{a_1 H_1 q_1^2}\frac{1}{1+z_2'}\left[q_1 z_2' + (q_1-1)\left(\sqrt{1+2q_1z_2'}-1\right)\right]. \tag{A.27}$$

To assign a value to the deceleration parameter q_0,

$$q_0 \equiv -\left(\frac{\ddot{a}a}{\dot{a}^2}\right)_{t=t_0}, \tag{A.28}$$

several assumptions must be made. The equation of state, which relates energy density ϵ and pressure P (not to be confused with the P in the Liske equations), is usually quite complicated. In cosmology, which deals primarily with dilute gasses, it takes the relatively simple form of $P = w\epsilon$ where w is a dimensionless number. The most important values of w in a cosmological sense are those for nonrelativistic gases, relativistic gases (e.g. photons), and dark energy,

$$w_{nonrel} \approx \frac{\langle v^2\rangle}{3c^2} \ll 1, \tag{A.29}$$

$$w_{rel} = \frac{1}{3}, \tag{A.30}$$

$$w_\Lambda \approx -1. \tag{A.31}$$

The WMAP 7 results (Jarosik et al., 2011) calculate $w_\Lambda = -1.12^{+0.42}_{-0.43}$ from WMAP alone and $w_\Lambda = -0.980 \pm 0.053$ if BAO and H_0 data are also included.

Combining the Friedmann acceleration equation,

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P)\,, \tag{A.32}$$

with the definition of q_0 and the trivial requirements that $\epsilon = \sum_w \epsilon_w$ and $P = \sum_w w\epsilon_w$, one can derive

$$q_0 = \frac{1}{2}\left(\frac{8\pi G}{3c^2H^2}\right)\sum_w \epsilon_w(1+3w). \tag{A.33}$$

The Friedmann equation in a flat Universe is

$$H(t)^2 = \frac{8\pi G}{3c^2}\epsilon(t). \tag{A.34}$$

If we take ϵ_c to be the critical energy density of the Universe, the combination of equations (A.33) and (A.34) yield

$$\begin{aligned} q_0 &= \frac{1}{2\epsilon_c}\sum_w \epsilon_w(1+3w) \\ &= \frac{1}{2}\sum_w \Omega_{w,0}(1+3w) \\ &= \Omega_{r,0} + \frac{1}{2}\Omega_{m,0} - \Omega_{\Lambda,0} \\ &= -0.55, \end{aligned} \tag{A.35}$$

Appendix B

Appendix B - Coordinate Transformation in the SDSS

The Sloan Digital Sky Survey uses a handful of different coordinate systems. Throughout my work I found it necessary to effectively convert between celestial and Cartesian coordinates. I include here a mapping between the two systems as defined within the SDSS database.

Transforming from celestial to Cartesian coordinates is straightforward,

$$\begin{aligned} c_x &= \cos(dec)\cos(\mathrm{RA}), \\ c_y &= \cos(dec)\sin(\mathrm{RA}), \\ c_z &= \sin(dec), \end{aligned} \tag{B.1}$$

where $RA \in [0, 2\pi)$ and $dec \in (-\pi/2, \pi/2)$.

Moving in the other direction takes greater delicacy because of wrap-around effects and divisions by zero. To maintain the same range of right ascension and declination, the following algorithm can be employed. Be careful that the inverse tangent is returned in degrees.

$$dec = \sin^{-1} c_z,$$

$$\mathrm{RA} = \begin{cases} \tan^{-1}(c_y/c_x) & \text{if } c_x > 0 \text{ and } c_y \geq 0 \\ \tan^{-1}(c_y/c_x) + 180 & \text{if } c_x < 0 \\ \tan^{-1}(c_y/c_x) + 360 & \text{if } c_x > 0 \text{ and } c_y < 0. \end{cases} \tag{B.2}$$

The operation of these right ascension equations depends upon the range of values your program returns for the arctangent function. For both SQL and MATLAB, this algorithm maintains the proper range of values in the SDSS database.

Special conditions apply in the event that $c_x = 0$. Where the right ascension is undefined, we set it to zero by default.

c_x	c_y	c_z	**RA**	**dec**
0	0	1	undefined	$90°$
0	0	-1	undefined	$-90°$
0	1	0	$90°$	$0°$
0	-1	0	$270°$	$0°$
0	> 0	$\neq 0$	$90°$	$\sin^{-1} c_z$
0	< 0	$\neq 0$	$270°$	$\sin^{-1} c_z$

Appendix C

Appendix C – SQL Queries

This appendix contains explicit SQL queries that can be used to reproduce the samples referenced within this dissertation. The first section contains scripts to extract the geometric properties of SDSS regions. The second section contains scripts that produce the Main Galaxy Sample. Background information on bit flags and K corrections is also provided.

C.1 SDSS Geometry

The queries that return the geometric properties for all SEGMENTs, PRIMARYs, PRIMARY SEGMENTs, TILEs and SECTORs are included here. For details about these regions, refer to §2.2.

SEGMENTs

The `RegionConvex` table identifies the SEGMENTs, and the `HalfSpace` table returns their constraint condition 4-vectors (n_x, n_y, n_z, c). A distribution of SEGMENT lengths is shown in Figure C.1.

```
SELECT *
FROM HalfSpace
WHERE regionID IN (SELECT regionID
                   FROM RegionConvex
                   WHERE type = 'SEGMENT')
```

PRIMARYs, CHUNKs

Like SEGMENTs, PRIMARYs (and CHUNKs) are defined through four constraint conditions stored in the `HalfSpace` table.

```
SELECT regionid, constraintid, x, y, z, c
FROM HalfSpace
WHERE regionID IN (SELECT regionID
                   FROM RegionConvex
                   WHERE type = 'CHUNK' OR type = 'PRIMARY')
```

PRIMARY SEGMENTs

Each of the 2052 SEGMENTs belongs to one of the 111 PRIMARYs. This query identifies which SEGMENTs are associated with which PRIMARYs.

```
SELECT regionid as SegmentID, s.chunkID, c.regionID as PrimaryID
FROM Segment s, Region r, Region c
WHERE r.type = 'SEGMENT' AND
      r.id = s.segmentID AND
      c.type = 'PRIMARY' AND
      s.chunkID = c.id
ORDER BY r.regionID
```

A point within the primary portion of a SEGMENT must meet the constraint conditions

of both the SEGMENT *and* its corresponding PRIMARY, for a total of eight constraint conditions in all. The SQL procedure listed below returns a table with 2052*8=16,416 rows.

```
INSERT INTO PrSegConstraints
SELECT r.regionid as RegionID, h.constraintID, h.x, h.y, h.z, h.c
FROM Region r, Segment s, Region c, HalfSpace h
WHERE r.type = 'SEGMENT' AND
      r.id = s.segmentID AND
      c.type = 'PRIMARY' AND
      c.id = s.chunkid AND
      h.regionid = r.regionid

INSERT INTO PrSegConstraints
SELECT rc.RegionID, h.constraintid, h.x, h.y, h.z, h.c
FROM(SELECT r.regionid as RegionID, s.chunkID
     FROM Region r, Segment s, Region c
     WHERE r.type='SEGMENT' AND
           r.id=s.segmentid AND
           c.type='PRIMARY' AND
           c.id=s.chunkid) rc, Region g, HalfSpace h
WHERE rc.chunkID = g.ID AND
      g.type = 'PRIMARY' AND
      g.regionID = h.regionid
```

The distributions of PRIMARY SEGMENT and SEGMENT lengths are plotted in Figure C.1. The figure shows that there are more PRIMARY SEGMENTs of shorter lengths and fewer of intermediate and long lengths. This verifies that the addition of PRIMARY constraints has shortened the SEGMENTs by cropping the areas that lie outside the SEGMENTs' PRIMARYs.

The speed of searches over PRIMARY SEGMENTs depends on the order in which constraint conditions are applied. Because SEGMENTs are smaller than PRIMARYs, fewer points will survive the former's constraints than the latter's. This suggests that searching over SEGMENTs first will reduce the number of required mathematical operations. As-

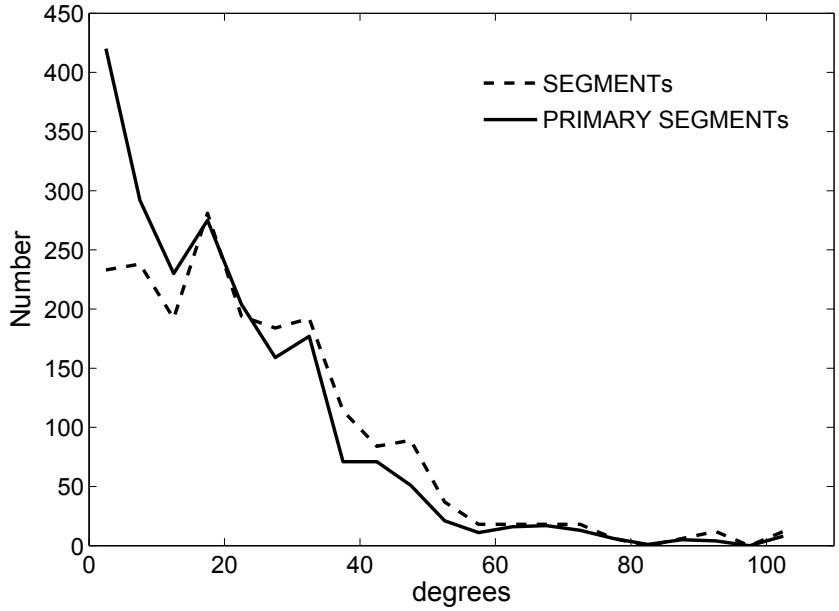

Figure C.1: Distribution of SEGMENT and PRIMARY SEGMENT lengths for DR6. Lengths are counted in bins of width 5°.

suming the constraint conditions are applied in the order they exist in the output table, a reverse ordering of the *constraintID*'s is preferable:

```
SELECT *
FROM PrSegConstraints
ORDER BY regionID, constraintID DESC
```

Such a sorting was used to measure the lengths of the regions in Figure C.1. High-density, uniformly distributed angular randoms were placed in the vicinity of the regions, then the constraints were applied. The length of each region was set equal to the largest angular separation among the points that survived the filtering.

This process revealed that 116 PRIMARY SEGMENTs have lengths equal or approximately equal to zero. Therefore, the number of degrees of freedom of the zero-points is not actually 2052, but something closer to 1936. Using 180 million full-sky angular randoms,

this filtering also revealed that approximately 20.13% lie within the union PRIMARY SEGMENTs. This corresponds to a footprint of about 8304.59 deg^2 .

TILEs

Each tile is circular and, thus, has only one constraint condition.

```
SELECT h.regionid, h.x, h.y, h.z, h.c
FROM Region r, HalfSpace h
WHERE r.type = 'TILE' AND
      r.regionid = h.regionid
```

The following query returns the SECTORs associated with each TILE. The section in green excludes SECTORs that are masks for a particular TILE. These masks are exclusively outside their associated TILE's boundaries and were only used in the past to serve as masks for other TILEs.

```
SELECT b.tile, b.regionID as sectorID, ns
FROM(SELECT tile, count(*) as ns
     FROM(SELECT *
          FROM Sector2Tile
          WHERE type = 'SECTOR' AND isMask=0),
     GROUP BY tile) a,
    (SELECT *
     FROM Sector2Tile
     WHERE type = 'SECTOR' AND isMask = 0) b
WHERE a.tile = b.tile
```

SECTORs

SECTORs are the complex intersections of multiple TILEs and tile masks, and can be unions of anywhere between 1 and 12 constraint conditions. To search over SECTORs, the user must first determine the number of associated convexes. The constraint conditions for

each convex are applied one at a time. Points that satisfy all of any convex's constraints lie within the SECTOR.

The following query returns all the information needed to determine whether a point lies within a SECTOR.

```
SELECT DISTINCT r.regionID, h.convexid, b.nc, h.x, h.y, h.z, h.c,
                e.nconvex
FROM Region r, HalfSpace h,
    (SELECT a.regionID, a.convexid, count(a.convexid) as nc
     FROM(SELECT DISTINCT r.regionID, h.convexid, h.x, h.y, h.z,
                          h.c
          FROM Region r, HalfSpace h
          WHERE r.type = 'SECTOR' AND r.regionid = h.regionid) a
     GROUP BY a.regionID, a.convexid) b,
    (SELECT d.regionID, count(d.regionID) as nconvex
     FROM(SELECT c.regionID, c.convexid, count(c.convexid) as nc
          FROM(SELECT DISTINCT r.regionID, h.convexid, h.x, h.y,
                               h.z, h.c
               FROM Region r, HalfSpace h
               WHERE r.type = 'SECTOR' AND
                     r.regionid = h.regionid) c
          GROUP BY c.regionID, c.convexid) d
     GROUP BY d.regionID) e
WHERE r.type = 'SECTOR' AND
      r.regionid = h.regionid AND
      r.regionid = b.regionid AND
      h.convexid = b.convexid AND
      e.regionid = h.regionid
ORDER BY r.regionID, e.nconvex, h.convexid, h.c DESC
```

C.2 Main Galaxy Sample

This section contains the scripts used to extract the MGS. The three samples whose queries are discussed here are: 1) pristine galaxies, 2) no-redshift objects, and 3) low-quality redshift objects. Drawing these targets from the CAS database requires the use of

bitwise arithmetic and K corrections. Explanations of both concepts are provided here.

The following code creates a full, unedited list of objects identified as MGS targets by virtue of their *PrimTarget* flag. The clustered index commands reorder the table to maximize search speeds over the fields *ObjID* and *SpecObjID*. Clustered indices were utilized for the other queries in this appendix, but are omitted elsewhere for brevity.

```
INSERT TEMP_3
SELECT p.objID, p.specObjID, p.primTarget, p.petroMag_r,
       p.extinction_r, p.cx, p.cy, p.cz, p.ra, p.dec
FROM PhotoPrimary p
WHERE (p.primtarget  448) != 0

CREATE CLUSTERED INDEX [i_DR6p3_ObjID] ON [dbo].[TEMP_3]
   ([ObjID] ASC) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [i_DR6p3_cxyz] ON [dbo].[TEMP_3]
   ([SpecObjID] ASC) ON [PRIMARY]
```

MGS targets have their *PrimTarget* flag set to at least one of the following:

64 = 'TARGET_GALAXY'

128 = 'TARGET_GALAXY_BIG' and/or

256 = 'TARGET_GALAXY_BRIGHT_CORE'.

Information in flags is stored in binary and manipulated through bitwise arithmetic. The bitwise summation operator & "adds" numbers in binary, generating 1 if two aligned digits are 1 and 0 otherwise.

Respectively, the three categories above are represented by setting the 7^{th}, 8^{th} and 9^{th} bits to 1, such that an object satisfying all 3 categories simultaneously has the flag 0001 1100 0000=448=256+128+64. Therefore, only a flag with a 1 in at least one of these

three positions qualifies as an MGS candidate. In SQL, these can be selected by requiring `(primtarget & 448)!= 0`.

Next, each MGS target is assigned a K correction derived through its spectral characteristics. In this query, the K corrections are stored in the table `KcorrDR6`:

```
INSERT TEMP_2
SELECT p.*, k.kr, k.k1r
FROM TEMP_3 p, KcorrDR6 k
WHERE p.specObjID = k.specObjID AND
      p.specObjID != 0

INSERT TEMP_2
SELECT p.*, -9999, -9999
FROM TEMP_3 p
WHERE p.specObjID NOT IN (SELECT specObjID FROM KcorrDR6) OR
      p.specObjID = 0
```

The K correction is defined by the manner in which it modifies the distance modulus of a galaxy at redshift z_i,

$$M(z_i) = m - (5 \log d_L(z_i) + 25) - k(z_i). \tag{C.1}$$

The reasoning behind this correction is that a photon emitted from a source at redshift z with frequency ν_e will be observed at a frequency ν_0 where $\nu_e = (1 + z)\nu_0$. Therefore, the luminosity through a restframe filter will differ from that through an emitted-frame filter as shown in Figure C.2.

The K correction shifts the luminosity of each galaxy into a common restframe using SED filters similar to those employed in calculating template-based photo-z's. Values of $k(z_i)$ for each galaxy are drawn from the closest non-negative linear combination of the

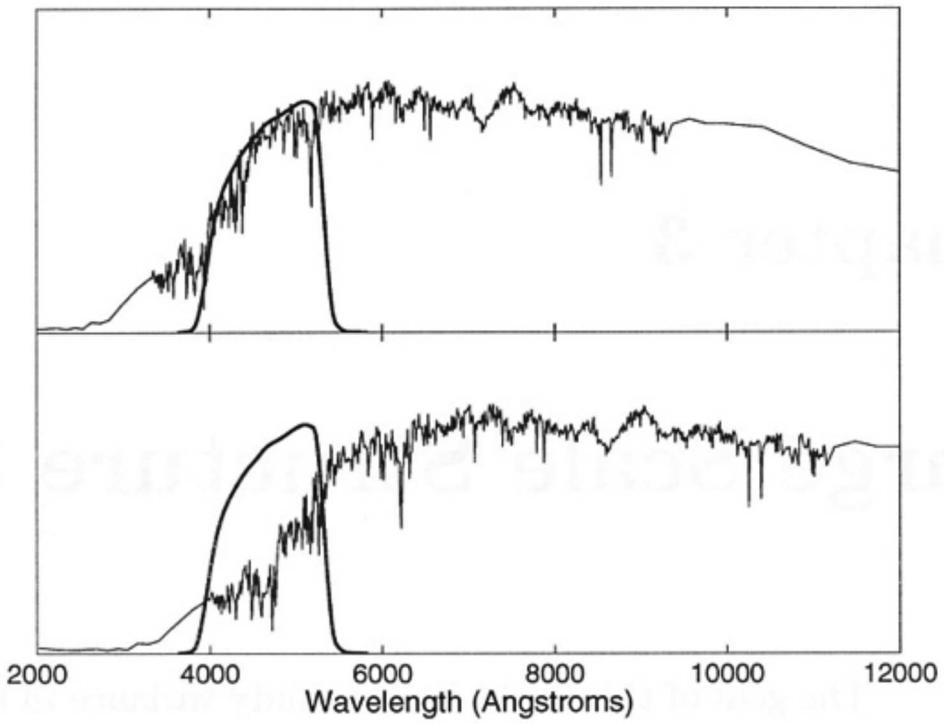

Figure C.2: Identical spectra (*jagged line*) measured in the rest frame (*top*) and the emitted-frame (*bottom*). A bandpass filter response centered on 5000Åis superimposed as a smooth curve. Adjustment of the distance modulus using a K corrections in equation (C.1) is needed before objects at different redshifts can be properly compared.

plates. This makes a galaxy's K correction just as much a function of its spectral type as its redshift.

The values we utilized in the `KcorrDR6` table were generated at JHU by Manuchehr Taghizadeh-Popp by adapting the earlier work of Tamas Budavári. This table and those for other data releases were only used internally and never released publically. For more on the principles behind the generation of these K corrections, see Hogg et al. (2002); Bruzual & Charlot (2003); Budavári et al. (2000); Csabai et al. (2000).

In the process of parameterizing the selection function (see §2.3.3.1) it became neces-

sary to provide a deterministic equation relating K corrections to redshift. However, due to the variable nature of galaxy spectra, objects at different redshifts can have the same K correction, and vice-versa.

Our solution was to define the averaged K correction, $k(z)$, as the Gaussian weighted average of $k(z_i)$ over the MGS pristine galaxies,

$$k(z) \equiv \frac{\sum_{i=1}^{N} w(z, z_i) k(z_i)}{\sum_{i=1}^{N} w(z, z_i)}, \tag{C.2}$$

where

$$w(z, z_i) = \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{z_i - z}{\sigma}\right)^2\right) & |z_i - z| \leq 0.01 \\ 0 & \text{otherwise} \end{cases}. \tag{C.3}$$

The total number of galaxies is N, and σ is a characteristic smoothing length chosen such that $3\sigma = 0.01$. (On the scale $\Delta z = 0.01$, $k(z)$ is approximately linear, so averaging within this range should be acceptable.) Unlike the MGS tables, which are populated with K corrections from DR6, we used the updated and improved K corrections of DR8 to determine $k(z)$. The results from both data releases were consistent.

The function $k(z)$ is subject to the variance of MGS galaxies used to generate it. It is preferable to use K correction values drawn from the best fit line displayed in Figure C.3. While a quadratic curve better fits the data, a linear response is expected since frequency scales with z, not z^2. Also, this linear fit better accommodates the empirically derived K

corrections for the 823 MGS pristine galaxies (assuming the redshift criterion is lifted) at $z > 0.3$.

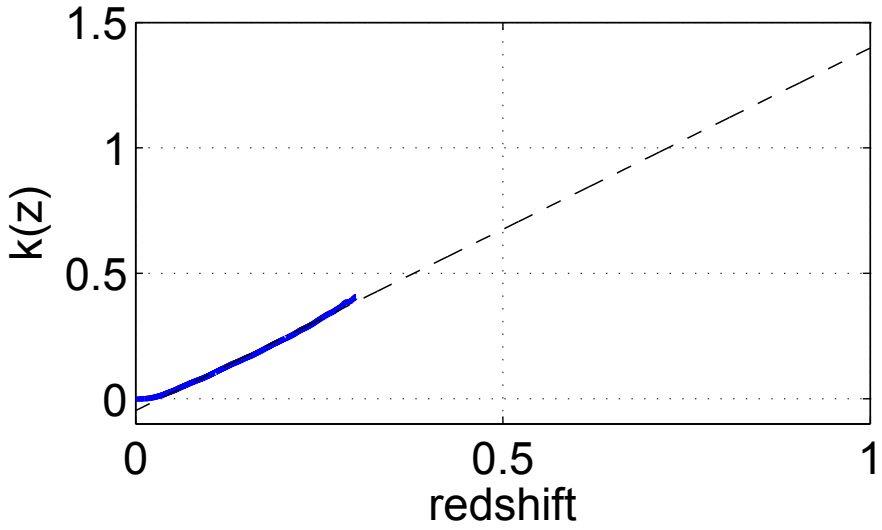

Figure C.3: Average r-band K corrections for DR8 MGS pristine galaxies as determined from using a Gaussian weighted average as derived from equations (C.2) and (C.3). The linear best-fit (*dashed*) line is fit from $k(z)$ using galaxies at $0.02 \leq z \leq 0.30$ where the spacing between $k(z)$ values is $\Delta z = 10^{-4}$.

The template photo-z's are drawn from the DR6 `Photoz` table. For the one target lacking these photometric redshifts, values of -9999 are placed in their stead.

```
INSERT TEMP_1
SELECT p.*, z.z, z.zErr
FROM TEMP_1 p, Photoz z
WHERE p.ObjID = z.ObjID

INSERT TEMP_1
SELECT p.*, -9999, -9999
FROM TEMP_2 p
WHERE p.ObjID NOT IN (SELECT ObjID FROM Photoz)
```

A similar query is used for the training-set (ANN) photo-z's. All objects missing CC2 photo-z's are also missing D1 photo-z's, and vice-versa.

```
INSERT DR6_PrimTarget448
SELECT p.*, z.photozcc2, z.photozerrcc2, z.photozd1, z.photozerrd1
FROM TEMP_1 p, Photoz2 z
WHERE p.ObjID = z.ObjID

INSERT DR6_PrimTarget448
SELECT p.*, -9999, -9999, -9999, -9999
FROM TEMP_1 p
WHERE p.ObjID NOT IN (SELECT ObjID FROM Photoz2)
```

The following query incorporates spectroscopic information to query the DR6 MGS pristine galaxies:

```
SELECT p.*, s.specObjID, s.zStatus, s.z, s.zconf, s.specClass
FROM DR6_PrimTarget448 p, SpecObj s
WHERE (p.petromag_r-p.extinction_r) > 15 AND
      p.specObjID != 0 AND
      p.specObjID = s.specObjID AND
      s.specClass = 2 AND
      s.z <= 0.22 AND
      s.z >= 0.02 AND
      s.zStatus NOT IN (0,1,10,2,5,8) AND
      s.zconf >= 0.9 AND
      (p.petroMag_r - p.extinction_r) - p.distmod - p.kr <= -17
```

This query extracts all data needed for the no-redshift objects table:

```
INSERT DR6_MGS_nospectra
SELECT p.*
FROM DR6_PrimTarget448 p
WHERE ((p.petromag_r-p.extinction_r) > 15 AND specObjID = 0) OR
      objID IN (SELECT pr.objID
                FROM DR6_primtarget448 pr, SpecObj s
                WHERE (pr.petromag_r-pr.extinction_r) > 15 AND
                      pr.specObjID != 0 AND
                      s.specObjID = pr.specObjID AND
                      s.zstatus IN (0,1,2))
```

Finally, this query yields the low-quality redshift objects sample. Any object with a redshift in the range $0 < z \leq 1$ is included.

```
DECLARE @h float,
```

```
        @BigMmax float;

SET @h = 0.70000000000000000;
SET @BigMmax = -17;

INSERT DR6_MGS_lowq
SELECT p.* s.z, s.zStatus, s.zconf, s.specClass
FROM DR6_PrimTarget448 p, SpecObj s
WHERE (pr.petromag_r-pr.extinction_r) > 15 AND
      p.specObjID != 0 AND
      p.specObjID = s.specObjID AND
      (s.zstatus IN (10,5,8) OR
       (s.zstatus NOT IN (0,1,10,2,5,8) AND s.zConf < 0.9)) AND
      s.z > 0 AND
      s.z <= 1
```

Appendix D

Appendix D – Power Spectra Derivations

In this Appendix we provide supporting materials for the power spectra explanations of §3.2. First, we summarize a conventional method of calculating power spectra — by representing each galaxy as a superposition of plane waves in Fourier space. Then we review how peculiar velocities can introduce anisotropies into the power spectrum, necessitating the use of redshift-space corrections.

D.1 Superposition of Plane Waves

Let $\delta^{(3)}$ represent the three-dimensional Dirac delta function. In a survey with N galaxies located at points $\boldsymbol{x}_i$, the number of galaxies at $\boldsymbol{x}$ is $\sum_{i=1}^{N} \delta^{(3)}(\boldsymbol{x} - \boldsymbol{x}_i)$. Without a weighting function $w(\boldsymbol{x}_i)$ to account for the drop in the number of observed galaxies as a function of depth in a magnitude limited survey, the count function places too much weight

on nearby galaxies. An improvement, $\tilde{\rho}(\boldsymbol{x})$, goes as $\sum_{i=1}^{N} w(\boldsymbol{x}_i)\,\delta^{(3)}(\boldsymbol{x}-\boldsymbol{x}_i)$ divided by $\sum_{i=1}^{N} w(\boldsymbol{x}_i)$ to normalize. The density symbol reflects that the Dirac delta function has units of inverse volume.

In Fourier space, $\tilde{\rho}(\boldsymbol{k}) = \Big(1/\sum_{i=1}^{N} w(\boldsymbol{x}_i)\Big) \sum_{i=1}^{N} w(\boldsymbol{x}_i) \int \delta^{(3)}(\boldsymbol{x}-\boldsymbol{x}_i)\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} d^3k$, which by virtue of the Dirac delta simplifies to $\tilde{\rho}(\boldsymbol{k}) = \Big(1/\sum_{i=1}^{N} w(\boldsymbol{x}_i)\Big) \sum_{i=1}^{N} w(\boldsymbol{x}_i) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_i}$. Subtracting the contribution from the window function of the survey $\tilde{W}(\boldsymbol{k})$ ($\tilde{W}(\boldsymbol{x}) = 0$ outside the survey and 1 inside) leaves only the structure of the data itself. Because of the normalizations, subtracting $\tilde{W}(\boldsymbol{k})$ also removes the DC component (i.e. the mean or $k = 0$ component) of the data.

Of course, the real measure of counts in space $\rho(\boldsymbol{x})$ is unknown. We are constrained by the window function of our survey such that $\tilde{\rho}(\boldsymbol{x}) = W(\boldsymbol{x})\rho(\boldsymbol{x}) = W(\boldsymbol{x})\rho_{bg}(1+\delta(\boldsymbol{x}))$. The Fourier transform is normalized such that the DC component vanishes. Accordingly, we take $\rho_{bg} = 1$ here so that $\tilde{\rho}(\boldsymbol{x}) = W(\boldsymbol{x}) + W(\boldsymbol{x})\delta(\boldsymbol{x})$. The Fourier transform of a product is a convolution such that $\tilde{\rho}(\boldsymbol{k}) = W(\boldsymbol{k}) + W(\boldsymbol{k}) * \delta(\boldsymbol{k}) = W(\boldsymbol{k}) + \tilde{\delta}(\boldsymbol{k})$. From this, the result cited in equation (3.20) emerges,

$$\tilde{\delta}(\boldsymbol{k}) = \frac{1}{\sum_j w(\boldsymbol{x}_j)} \sum_j w(\boldsymbol{x}_j) e^{i\boldsymbol{k}\cdot\boldsymbol{x}_j} - \tilde{W}(\boldsymbol{k}). \tag{D.1}$$

The same function evaluated continuously is

$$\tilde{\delta}(\boldsymbol{k}) = \int d^3k'\, \delta(\boldsymbol{k}-\boldsymbol{k}')\, W(\boldsymbol{k}'). \tag{D.2}$$

The convolution has two effects on $\delta(\boldsymbol{k})$. It changes the shape (sharp features become broadened) and the norm. We cannot correct the shape but we can renormalize to fix the norm.

The power spectrum comes from the product of the density with its complex conjugate. A manipulation of equation (D.1) provides a sense of how the shot noise term interjects itself in measures of power,

$$\tilde{\rho}(\boldsymbol{k})\tilde{\rho}(\boldsymbol{k})^* = \frac{1}{\left(\sum_{i=1}^{N} w(\boldsymbol{x}_i)\right)^2} \sum_{i,j} w(\boldsymbol{x}_i) w(\boldsymbol{x}_j) e^{-i\,\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{D.3}$$

This can be split into a sum when $i = j$ and one when $i \neq j$,

$$\tilde{\rho}(\boldsymbol{k})\tilde{\rho}(\boldsymbol{k})^* = \frac{1}{\left(\sum_{i=1}^{N} w(\boldsymbol{x}_i)\right)^2} \left[\sum_{i,j} w(\boldsymbol{x}_i)^2 \cdot 1 + \sum_{i,j} w(\boldsymbol{x}_i) w(\boldsymbol{x}_j) e^{-i\,\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)} \right]. \tag{D.4}$$

The first term within the brackets is a shot noise term that ought to be subtracted out when reporting structure on its own.

One weight candidate, that of FKP94, was introduced in equation (3.21). Another, from Percival et al. (2007), argues in favor of a bias-dependent weighting where $\bar{P}(k)$ is an unbiased power spectrum estimate,

$$w(\boldsymbol{r}, b') = \frac{(b')^2(\boldsymbol{r})\bar{P}(k)}{1 + \int db\, \langle S(\boldsymbol{r}, b)\rangle b^2 \bar{P}(k)}. \tag{D.5}$$

However, both forms fail to model an effect that plagues the MGS — anisotropic angular completeness. Just as the selection function appropriately weights galaxies to account for their lower detection rate at high z, so will the $a(\boldsymbol{x})$ upweight galaxies in regions with low angular completeness. Therefore we propose the following improvement to the simple FKP weighting of equation (3.21),

$$w'(\boldsymbol{x}_j) = \frac{1}{S(r_j)a(\hat{\boldsymbol{x}}_j)}. \tag{D.6}$$

We should note that any weighting scheme is somewhat arbitrary. In equation (D.6) we have used the inverse of the selection function weighted by the angular completeness. FKP94 incorporates the inverse variance (i.e. power) of the quantity they are trying to measure (i.e. density). It can be argued that one must take two effects into consideration when measuring power on some scale k — the sparsity of the shot noise (from sampling) and the "natural" variance of the power. FKP94 arrives at equation (3.21) by solving for the optimal weighting under the assumption of Gaussian fluctuations.

Either way, the window function is calculated by populating the survey footprint with high density synthetic points at positions $\boldsymbol{x}_j$ and evaluating

$$\tilde{W}(\boldsymbol{k}) = \frac{\int d^3x\, W(\boldsymbol{x}) e^{i\boldsymbol{k}\cdot\boldsymbol{x}}}{\int d^3x\, W(\boldsymbol{x})} \cong \frac{\sum_j W(x_j)\cdot e^{i\boldsymbol{k}\cdot\boldsymbol{x}_j}}{\sum_j W(x_j)}. \tag{D.7}$$

The estimate of the power spectrum $\hat{P}(\boldsymbol{k})$ is then given as

$$\hat{P}(\boldsymbol{k}) = \frac{|\tilde{\delta}(\boldsymbol{k})|^2 - \sum_j w'^2(\boldsymbol{x}_j) / \left[\sum_j w'(\boldsymbol{x}_j)\right]^2}{[1/(2\pi)^3] \int d^3k' \, |\tilde{W}(\boldsymbol{k}')|^2}. \tag{D.8}$$

In the absence of weights, the second term in the numerator equals $1/N$, or the shot noise when N is the total number of galaxies.

D.2 Redshift-Space Distortions

In a linear-regime galaxy cluster where perturbations are assumed small, a real-space overdensity $\delta^{(r)}$ will induce radial peculiar velocities v_r along the line-of-sight. The continuity equation $\dot{\delta}_k + (ik_\alpha v_\alpha(\boldsymbol{k}))/a = 0$ relates the two for each component α once it is evolved in time through the growth factor,

$$v_r(\boldsymbol{k}) = \frac{ik_r}{k^2} H_0 \, \alpha_0 \, \beta \, \delta^{(r)}(\boldsymbol{k}), \tag{D.9}$$

where

$$\beta = \frac{f}{b} \cong \frac{\Omega_m^{0.6}}{b}. \tag{D.10}$$

The radial component of the wave vector in Fourier space is $k_r = |\boldsymbol{k}|\hat{k} \cdot \hat{r} = k\mu$. The scale factor f is related to the growth of structure in the Universe. For a wide range of cosmological parameters $f = \Omega_m^{0.6}$.

We set this aside for the moment and consider the relationship between real-space and

redshift-space. A galaxy appears at a redshift distance $s = \boldsymbol{s} \cdot \hat{r}$ as a result of its Hubble flow-induced real-space position $r = \boldsymbol{r} \cdot \hat{r}$ and its peculiar velocity v_r,

$$s = r + \frac{v_r}{H_0}. \tag{D.11}$$

Kaiser (1987) recognized that the number of galaxies within a region will be the same regardless of whether one uses real-space or redshift-space coordinates, as the conversion merely changes the shape of the region, not the amount of mass therein,

$$\rho(\boldsymbol{s})\, d^3\boldsymbol{s} = \rho(\boldsymbol{r})\, d^3\boldsymbol{r}. \tag{D.12}$$

Redshift-space distortions do not affect angular directions, so a coordinate transformation through the Jacobian reveals

$$\rho^{(s)} \cong \rho^{(r)} \left(1 + \frac{\partial v_r}{H_0\, \partial r}\right)^{-1} \cong \rho^{(r)} \left(1 - \frac{\partial v_r}{H_0\, \partial r}\right). \tag{D.13}$$

The recession velocities in the high-redshift Universe grow larger while the peculiar velocities stay roughly the same. Therefore the derivative $\partial v_r / \partial r$ asymptotically approaches zero and justifies the binomial approximation here.

We approximate that the background densities in real-space and redshift-space are the same, convert the densities to overdensities, and expand to first order,

$$1 + \delta^{(s)} = \left(1 + \delta^{(r)}\right)\left(1 - \frac{\partial v_r}{H_0\,\partial r}\right) \cong 1 + \delta^{(r)} - \frac{\partial v_r}{H_0\,\partial r}, \tag{D.14}$$

$$\delta^{(s)} = \delta^{(r)} - \frac{\partial v_r}{H_0\,\partial r}. \tag{D.15}$$

In Fourier space $\partial/\partial r \sim ik\mu$, so

$$\delta^{(s)}(\boldsymbol{k}) = \delta^{(r)}(\boldsymbol{k}) - \frac{ik\mu}{H_0} v_r(\boldsymbol{k}) = \delta^{(r)}(\boldsymbol{k})\left(1 + \beta\mu^2\right). \tag{D.16}$$

Consequently, the spherically symmetric power spectrum in real-space, $P(k)$, is modulated by the factor β and becomes anisotropic via the cosine between the wave vector and the line-of-sight μ,

$$P^{(s)}(\boldsymbol{k}) = P(k)\left(1 + \beta\mu^2\right)^2. \tag{D.17}$$

When $b = 1.2$ and $\Omega_m = 0.3$, we find $\beta = 0.405$. Along the line-of-sight, $\mu = 1$ and $(1 + \beta\mu^2)^2 \approx 2$. If $\boldsymbol{k}$ is perpendicular to the line-of-sight, then $(1 + \beta\mu^2)^2 = 1$. This suggests that the MGS power spectrum in redshift-space will be deformed like a football relative to real-space. The power approximately doubles, or "elongates", along the line-of-sight while the perpendicular direction remains unaffected.

Appendix E

Appendix E - Cell/Region Intersections and Angular Randoms Theory

This appendix describes how to determine the volume of each cell intersected by regions such as PRIMARY SEGMENTs and SECTORs. First, we explain how to generate angular random points, both over the full sky and in limited areas, for use in Monte Carlo simulations. We derive the weight each point receives relative to the sphere it intersects. Then we provide guidelines for how to search over regions in order to calculate the volume intersections as efficiently as possible. We conclude by justifying the criterion that every sphere have at least 62% of its volume inside the spectroscopic footprint.

We define an *angular random point* to be a randomly selected point on the unit sphere. It can be represented using three Cartesian coordinates or through its two degrees of freedom — an azimuthal coordinate $\theta \in (-\pi, \pi)$ and an altitudinal coordinate $\phi \in (0, \pi)$. The

APPENDIX E. APPENDIX E - CELL/REGION INTERSECTIONS AND ANGULAR RANDOMS THEORY

Cartesian representation is useful for generating full-sky angular randoms while the polar representation is better suited for generating randoms within angular limits.

To generate full sky randoms we take $u = \cos\varphi \in [-1, 1]$ and $\gamma \in [0, 2\pi)$ to both be uniformly distributed random variables. It follows that the Cartesian coordinates

$$\begin{aligned} x &= \sqrt{1-u^2}\cos\gamma, \\ y &= \sqrt{1-u^2}\sin\gamma, \\ z &= u, \end{aligned} \tag{E.1}$$

are distributed with a uniform density per unit area on the unit sphere. On the order of 10 million angular randoms can be generated per second, fast enough so that this is not a bottleneck.

Localized angular randoms are generated between limits such that $RA \in [RA_i, RA_f]$ and $dec \in [dec_i, dec_f]$. Right ascension can be drawn as a uniform random variable within the given range. Declination, however, must be drawn using its probability density function, then integrating to get a cumulative distribution function (CDF). CDF's have a range $[0, 1]$ and can be used to reverse map a uniform random variable to the declination associated with its value.

To see how this works, consider an infinitesimal area of latitude on the unit sphere $dA = 2\pi\cos\theta\, dr$. Both the area element dA and the distribution function for altitude angle

are proportional to $\cos\theta$. Using a basic distribution function $p(\theta) = A\cos\theta$ for declination, we require the following to limit an angular random to $\theta \in (\theta_i, \theta_f)$,

$$\int p(\theta)\, d\theta = \int_{\theta_i}^{\theta_f} A\cos\theta\, d\theta = A(\sin\theta_f - \sin\theta_i) = 1. \tag{E.2}$$

Upon normalizing, $p(\theta) = \cos\theta\,(\sin\theta_f - \sin\theta_i)^{-1}$. Integrate between the declination limits to calculate the CDF,

$$F(\theta) = \int_{\theta_i}^{\theta} \frac{\cos\theta'}{\sin\theta_f - \sin\theta_i}\, d\theta' = \frac{\sin\theta - \sin\theta_i}{\sin\theta_f - \sin\theta_i}. \tag{E.3}$$

Treat $U = F(\theta)$ as a uniform random variable and solve for declination,

$$dec = \sin^{-1}(\sin dec_i + (\sin dec_f - \sin dec_i)\, U)\,. \tag{E.4}$$

Angular randoms points are useful insofar as they enable Monte Carlo calculations of the intersection volumes of cells and SDSS regions. A region intersecting a spherical cell's center will occupy more volume than one grazing its edge. When considered in aggregate, uniformly distributed rays passing through these regions can be proxies for volume if each ray is weighted by an amount proportional to its penetration distance l through the sphere.

For instance, if we let w_c equal the sum of all weights passing through a cell, and let w_{r_i} equal the sum of weights passing through region i, then the percent volume of the cell intersected by region i will equal the ratio w_{r_i}/w_c.

Computing a ray's penetration distance is aided by the observation that any line-of-

sight through a sphere can be made to pass through the plane of one of its great circles. Put another way, all line-of-sight chords pass through a circle with a radius equal to that of the sphere itself.

Consider the geometry of Figure E.1. We redefine θ to be the angle between a random ray of direction $\hat{\boldsymbol{x}}$ and the center of the sphere $\hat{\boldsymbol{n}}$ such that $c \equiv \cos\theta = \hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{x}}$. A chord's perpendicular bisector intersects the center of a circle, therefore,

$$t^2 + \frac{l^2}{4} = r^2. \tag{E.5}$$

The distance d to the cell's center obeys the expression $d^2 \sin^2\theta = t^2$. Combining with equation (E.5) yields $l = 2\sqrt{r^2 - d^2 \sin^2\theta}$. Using the identity $\sin^2\theta = 1 - \cos^2\theta = 1 - c^2$ results in the final form of the chord length,

$$l = 2\sqrt{r^2 - \chi^2(1 - c^2)}. \tag{E.6}$$

Normalizing such that a ray passing through the cell's center is given a weight of 1,

$$w = \frac{l}{2r} = \sqrt{1 - \left(\frac{\chi}{r}\right)^2 (1 - c^2)}. \tag{E.7}$$

In the case of probabilistic smearing it is necessary to know the depths at which chords enter and exit cells. Setting the origin to the cell's center, the equations for the line-of-sight and cell boundary are respectively, $y = x\cot\theta - d$ and $x^2 + y^2 = r^2$. Solving simultaneously, the y-components of the intersection points are $y = -d\sin^2\theta \pm$

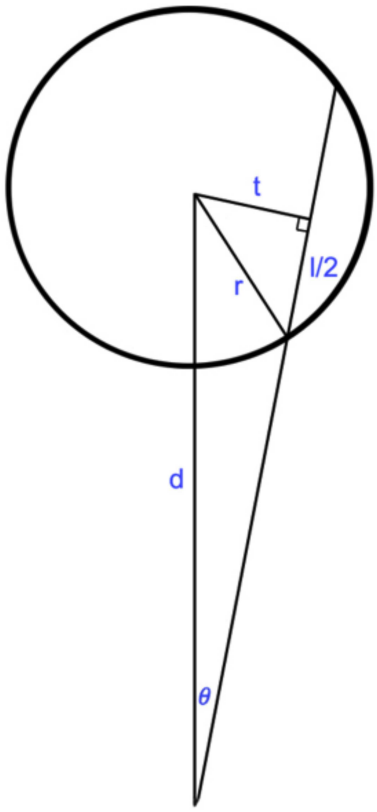

Figure E.1: View of angular random weighting geometry. The circle represents the great circle of the spherical cell through which a ray passes. The weight given to any chord equals the ratio of the chord length l to the diameter $2r$.

$\cos\theta\sqrt{r^2 - d^2\sin^2\theta}$, or

$$y_{\pm} = d\big(c^2 - 1\big) \pm c\sqrt{r^2 + d^2(c^2 - 1)}. \tag{E.8}$$

Shifting the origin back to the observer, the depths χ_l and χ_u at which the chord enters and exits the cell are

$$\chi_l = c^2\chi - c\sqrt{r^2 + \chi^2(c^2 - 1)},$$
$$\chi_u = c^2\chi + c\sqrt{r^2 + \chi^2(c^2 - 1)}. \tag{E.9}$$

Performing Monte Carlo volume estimations over all cells and regions can be computationally expensive, so it is important to order one's operations smartly. The first step is using equations (E.1) to populate the entire unit sphere with angular randoms. The circular projections of the most distant R7 cells occupy about 6×10^{-5} of the entire sky. A total of $N = n/6 \times 10^{-5}$ angular randoms ensures an average of at least n randoms per cell, or about 1 million total when $n = 50$. An application of the PRIMARY SEGMENT constraints filters the angular randoms so that only those within the photometric footprint remain.

Using the HCP, *full-sky cells* are positioned within a spherical volume of radius $z = 0.22$ for R7 and R11, or $z = 0.30$ for R16. Each cell's angular radius defines a circular constraint condition similar to those of TILEs. Any cell that contains one of the filtered angular randoms is deemed a "filtered cell" and remains a candidate for the final cell set. If the density of angular randoms is too low, cells that barely reach into the footprint may be incorrectly discarded. However, since our analysis only utilizes a cell if a majority of its volume lies within the footprint, a few false negatives are safely ignored.

After this initial filtering, about one-quarter of the full-sky cells remain as filtered cells.

Each must be studied individually to determine what fraction of its volume lies within each region. The regions we investigate include SECTORs — to determine what fraction lies within the spectroscopic footprint — and PRIMARY SEGMENTs — to determine both the fraction within the photometric footprint and to measure the effect each photometric offset will have on galaxy number count.

Full-sky angular randoms are filtered through each region's constraint conditions. The randoms with the minimum and maxium RA's and declinations are used to establish the boundaries of each region. These boundaries are compared against the boundaries of the filtered cells such that each cell becomes associated with a more limited set of regions that might intersect it. On average, about 65 PRIMARY SEGMENTs and 100 SECTORs intersect each cell in this way. This reduces the number of regions that must be searched over per cell by 1 to 2 orders of magnitude.

Using equation (E.4), high density random points are generated within each sphere. Equation (E.7) assigns weights to each point. Then w_{r_i}/w_c is used to calculate the intersection volume.

We stipulate that a cell is eligible for the final sample only if the fraction of its volume within the spectroscopic footprint β_{SPEC} is sufficiently large. We quantify this minimum volume by requiring that at least half the objects expected in each cell be pristine galaxies with high quality redshifts.

Assume we have a cell with a fraction of its volume β_{SPEC} inside the spectroscopic footprint. Further assume there are p pristine galaxies within that volume. If we assume all

three target types are represented proportionally to their overall number in the spectroscopic footprint, then

$$\begin{aligned} \langle n_i \rangle &= f_n p, \\ \langle l \rangle &= f_l p. \end{aligned} \tag{E.10}$$

where $\langle n_i \rangle$ and $\langle l \rangle$ are respectively the expected number of no-redshift and low-quality objects in the volume. The proportionality factors can be taken from the data where $f_n = 0.193$ and $f_l = 0.040$.

The cell's volume outside the spectroscopic footprint is a fraction $(1 - \beta_{SPEC}) / \beta_{SPEC}$ of that inside. Assuming a uniform density of targets overall, the number expected outside is

$$\langle n_o \rangle = \frac{(1 - \beta_{SPEC})}{\beta_{SPEC}} (p + \langle n_i \rangle + \langle l \rangle) . \tag{E.11}$$

It is reasonable to approximate that all of these are no-redshift targets. To satisfy the condition, there should be at least as many pristine targets as non-pristine targets,

$$p \geq \langle n_o \rangle + \langle n_i \rangle + \langle l \rangle. \tag{E.12}$$

Solving,

$$\beta_{SPEC} \geq \frac{1}{2}(1 + f_n + f_l)\,. \tag{E.13}$$

For DR6, this sets a threshold of $\beta_{SPEC} \cong 0.62$.

Appendix F

Appendix F – Shot Noise for Overlapping Cells

The shot noise covariance matrix $\mathbf{\Sigma}_\zeta$ is diagonal as long as the spheres do not overlap. If they do, the matrix must be adjusted to account for the new cross-correlations and nonzero diagonal elements. This appendix provides that derivation.

Consider two overlapping spherical cells of equal volume. Let d_i equal the number count of galaxies exclusively within cell i, d_j equal the number count of galaxies exclusively within cell j, and let d_{ij} equal the number count of galaxies in the overlap between cells i and j. The total number of galaxies in cell i will be $n_i = d_i + d_{ij}$ and the total number of galaxies in cell j will be $n_j = d_j + d_{ij}$.

We are interested in the expected value of the product of counts in cells, $\langle n_i n_j \rangle$. Let n equal the average number density of objects in each cell, V equal the volume of each cell,

and O_{ij} equal the overlap volume. Then, the expected number of objects in each cell in total, cell i exclusively, cell j exclusively, and the overlap between them are respectively,

$$\begin{aligned}
\lambda &= nV, \\
\lambda_i &= n(V - O_{ij}), \\
\lambda_j &= n(V - O_{ij}), \\
\lambda_{ij} &= nO_{ij}.
\end{aligned} \tag{F.1}$$

If two overlapping spheres of radius R have centers separated by a distance $s < 2R$ then,

$$O_{ij} = \frac{\pi}{12}\left(16R^3 - 12sR^2 + s^3\right). \tag{F.2}$$

We have assumed that the number densities of objects in two overlapping cells are the same. While this is unlikely to be strictly true, as long as the selection function doesn't change rapidly between cells, the approximation is sufficient. Evaluating the expected value of the product,

$$n_i n_j = d_i d_j + d_i d_{ij} + d_{ij} d_j + d_{ij}^2, \tag{F.3}$$

$$\langle n_i n_j \rangle = \langle d_i d_j \rangle + \langle d_i d_{ij} \rangle + \langle d_{ij} d_j \rangle + \langle d_{ij}^2 \rangle. \tag{F.4}$$

Since the counts in all three regions are independent of one another,

$$\begin{aligned}\langle n_i n_j \rangle &= \langle d_i \rangle \langle d_j \rangle + \langle d_i \rangle \langle d_{ij} \rangle + \langle d_{ij} \rangle \langle d_j \rangle + \langle d_{ij}^2 \rangle \\ &= \lambda_i \lambda_j + \lambda_i \lambda_{ij} + \lambda_{ij} \lambda_j + \lambda_{ij}^2 + \lambda_{ij}. \end{aligned} \tag{F.5}$$

The final two terms in equation (F.5) follow since d_{ij} is a Poisson distributed random variable for which $\mathrm{Var}(d_{ij}) = \lambda_{ij}$. Factoring equation (F.5),

$$\begin{aligned}\langle n_i n_j \rangle &= (\lambda_i + \lambda_{ij})(\lambda_j + \lambda_{ij}) + \lambda_{ij} \\ &= (n(V - O_{ij}) + nO_{ij})(n(V - O_{ij}) + nO_{ij}) + \lambda_{ij} \\ &= \lambda^2 + \lambda_{ij} = \langle n_i \rangle \langle n_j \rangle + \lambda_{ij}. \end{aligned} \tag{F.6}$$

Equation (F.6) evaluates to the following when the two cells are separate, when they completely overlap, and when they partially overlap,

$$\langle n_i n_j \rangle = \begin{cases} \langle n_i \rangle \langle n_j \rangle & \text{no overlap} \\ \langle n_i \rangle \langle n_j \rangle + \langle n_i \rangle & \text{full overlap} \\ \langle n_i \rangle \langle n_j \rangle + \lambda_{ij} & \text{partial overlap} \end{cases} . \tag{F.7}$$

Shot noise between nonoverlapping cells must be independent, therefore $\langle n_i n_j \rangle = \langle n_i \rangle \langle n_j \rangle$.

The fully overlapping case has an additional term $\langle n_i \rangle = \langle n_j \rangle$, while the partially overlapping case has an extra term of

$$\lambda_{ij} = n_{ij} = nV\left(\frac{O_{ij}}{V}\right) = \langle n_i \rangle \left(\frac{O_{ij}}{V}\right) = \langle n_j \rangle \left(\frac{O_{ij}}{V}\right). \tag{F.8}$$

This suggests that the overdensity correlation matrix in equation (3.10) can be modified in the following manner,

$$R_{ij} = \xi_{ij} + \left(\frac{O_{ij}}{V}\right)\frac{1}{\langle n_i \rangle} + \frac{\epsilon_{ij}}{\langle n_i \rangle \langle n_j \rangle}. \tag{F.9}$$

Note that in the case of nonoverlapping cells, the correlation matrix in equation (F.9) resumes the form first introduced in equation (3.10).

Appendix G

Appendix G – Alternative Methods

In the spirit of not letting negative results go to waste, the methods described in this section represent a suite of techniques that attempted, but failed, to adequately remove shot and systematic noise while retaining a quality signal. While these methods proved inappropriate for the overdensity cleansing problem, they should certainly be considered viable candidates for other classes of problems. The limitations and possible applications of these methods are described in the sections that follow.

All of these methods germinate from the observation that the zero-point noise has relatively few degrees of freedom. We reported that there are 2052 PRIMARY SEGMENTs defined in the DR6 database, and only 1890 of those have non-zero areas. In comparison, the number of degrees of freedom of the signal equals the dimensionality of discretization (i.e. the number of cells). As the number of cells increases, the zero-point noise is relegated to a diminishing fraction of the problem's overall dimensionality.

This reduction in dimensionality offers attractive opportunities. By localizing the zero-point noise in such a small number of dimensions, it should be possible to execute noise reduction techniques there without affecting the data residing along the vast majority of dimensions. One idea is to remove all data (e.g. signal and noise) that lie along the principle components of the noise, a process referred to as *deprojection*. We tried this in a number of different ways including deprojecting all 2052 dimensions and deprojecting a dimension only if doing so removed a greater fraction of the noise than the signal.

To analogize, our noise is a like a cancer that has spread throughout the body. By rotating the problem into the proper coordinate system, we reorganize that body so that the cancer becomes a localized tumor surrounded by healthy tissue (e.g. signal). Deprojection is akin to taking a cleaver to the tumor. In a swipe or two you can remove the tumor entirely but take an unacceptable amount of healthy tissue along with it. We found that deprojection could maximize the S/N ratio (where N is limited to zero-point noise), but the signal estimate it left failed to adequately represent the truth.

If the principle components of the signal and noise were non-overlapping, this method should have worked admirably. But as shown in Figures 4.6 and 4.14, the lowest order components tend to contain the largest scale structures. Consequently, the most important noise eigenmodes preferentially intersected the most important signal eigenmodes and the crosstalk between the spaces proved fatal to this approach.

We also attempted a collection of χ^2 minimizations in which we minimized the difference between the raw data vector $\boldsymbol{\delta}$ and a reconstructed signal plus noise model. Using

"truncated expansion" we found the linear combination of signal and noise modes that most closely approached $\boldsymbol{\delta}$. We also tried replacing the truncated noise with a best-fit set of $\Delta \boldsymbol{m}$ parameters through $\hat{\boldsymbol{\eta}} = \mathbf{A} \cdot \Delta \boldsymbol{m}$. In both cases, the cross talk remained too high.

Finally, following the method of Everson & Sirovich (1995), we tried a technique we termed "gappy reconstruction." This technique has been commonly used to restore images in which pixels are missing. Provided the complete image can be represented in an eigenbasis of lower rank than the number of pixels, the missing pixels are often able to be reconstructed.

In our case, we transformed our data vector into noise space and zeroed out the principle noise components, effectively creating gaps in the data. Using the signal eigenmodes we attempted to "fill in" those gaps with the algorithm's best guess for the signal that ought to have lied therein. Again, we encountered the familiar cross talk problem. We found that every successive noise element we zeroed out took with it more signal information than Everson Sirovich's method was able to reconstruct with the remaining information.

The moral of this story is that deprojection should be utilized only if there is minimal overlap between the principle signal and noise components. Otherwise, eliminating noise removes more signal than is acceptable — a fatal problem especially when the magnitude of the noise is small.

In the following sections, we outline the mathematics behinds the methods just described. We begin with a signal-to-noise maximization method whereby noise modes are deprojected only if doing so increases S/N. Next, we present the method of truncated re-

construction, which attempts to recreate signal and noise using a limited subset of their respective eigenmodes. Finally, we develop the theory behind gappy reconstruction, which uses signal covariance information to restore signal lost during the deprojection of principle noise modes.

G.1 Signal-to-Noise Maximization Method

In the signal-to-noise maximization method, noise modes are removed (i.e. the dimension is deprojected entirely) if doing so increases the signal-to-noise ratio. For simplicity, we shall lump the signal and shot noise together into *density modes*. Both the density modes and the zero-point noise modes have total variances quantified by the sums of their eigenvalues. The fractional variance captured by any one mode is the ratio of its eigenvalue to the total variance. However, since each noise mode can overlap with several density signal modes, it is important to quantify the fraction of the signal removed in this manner as well.

The true overdensity vector $\boldsymbol{\delta}$ may be expressed as a linear combination of N density modes $\hat{\boldsymbol{d}}_i$ and K zero-point noise modes $\hat{\boldsymbol{u}}_i$,

$$\boldsymbol{\delta} = \sum_{i=1}^{N} a_i \hat{\boldsymbol{d}}_i + \sum_{j=1}^{K} \mathrm{t}_j \hat{\boldsymbol{u}}_j. \tag{G.1}$$

In our case, the $K = 1890$ zero-point modes are divided into two mutually exclusive sets — $\{\hat{\boldsymbol{u}}'\}$, which contains the K' noise modes to be retained, and $\{\hat{\boldsymbol{u}}''\}$, which contains

the $K - K'$ noise modes to be deprojected. The estimated signal is then expressed as

$$\hat{\boldsymbol{\delta}}_\kappa = \sum_{i=1}^{N} a_i \hat{\boldsymbol{d}}_i + \sum_{j=1}^{K'} \mathfrak{t}_j \hat{\boldsymbol{u}}'_j - \sum_{k=1}^{K-K'} \mathfrak{t}_k \hat{\boldsymbol{u}}''_k. \tag{G.2}$$

Each noise mode $\hat{\boldsymbol{u}}_i$ overlaps with each density mode to an extent $\nu_{ij} = \hat{\boldsymbol{u}}_i^T$, where $\sum_{j=1}^{N} \nu_{ij}^2 = 1$. Furthermore, the fractional variance present in the i^{th} noise mode can be expressed as $\lambda_i^{(\eta)} / \sum_{j=1}^{K} \lambda_j^{(\eta)}$. The deprojection of $\hat{\boldsymbol{u}}_i$ will cause a fractional decrease in the signal equal to

$$\frac{\sum_{j=1}^{N} \nu_{ij}^2 \lambda_j^{(D)}}{\sum_{j=1}^{N} \lambda_j^{(D)}}. \tag{G.3}$$

The signal-to-noise maximization method dictates that the i^{th} noise mode $\hat{\boldsymbol{u}}_i$ should be removed from the data set provided

$$\frac{\lambda_i^{(\eta)}}{\sum_{j=1}^{K} \lambda_j^{(\eta)}} > \frac{\sum_{j=1}^{N} \nu_{ij}^2 \lambda_j^{(D)}}{\sum_{j=1}^{N} \lambda_j^{(D)}}. \tag{G.4}$$

For our problem, this condition is met for approximately the first ~ 700 contiguous noise modes, plus or minus a couple hundred depending on the size of the cells. According to numerous metrics, however, maximization of S/N did not improve the quality of the recovered signal on balance.

G.2 Truncated Reconstruction

An N-dimensional clustering overdensity vector $\boldsymbol{\delta}_\kappa$ can be approximated as a linear combination of $p < N$ principle signal components $\hat{\boldsymbol{z}}_i$. The k^{th} element of that vector would be

$$\boldsymbol{\delta}_\kappa[k] \approx \sum_{i=I_z[1]}^{I_z[p]} S_i \hat{\boldsymbol{z}}_i[k], \tag{G.5}$$

where I_z is an index set containing indices of the p signal dimensions expanded over. Likewise, the zero-point noise can be expanded over $q < K$ zero-point noise modes,

$$\boldsymbol{\delta}_\eta[k] \approx \sum_{j=I_u[1]}^{I_u[q]} \mathfrak{t}_j \hat{\boldsymbol{u}}_j[k], \tag{G.6}$$

where I_u is an index set containing indices of the q zero-point dimensions expanded over. Ignoring shot noise, the k^{th} component of a raw overdensity vector $\boldsymbol{\delta}$ can therefore be approximated

$$\boldsymbol{\delta}[k] \approx \sum_{i=I_z[1]}^{I_z[p]} S_i \hat{\boldsymbol{z}}_i[k] + \sum_{j=I_u[1]}^{I_u[q]} \mathfrak{t}_j \hat{\boldsymbol{u}}_j[k]. \tag{G.7}$$

Our goal is to find the index sets I_z and I_u that best approximate $\boldsymbol{\delta}$. This can be approached as a least squares minimization problem in which we seek to evaluate the following expression,

$$\underset{I_z,I_u}{\arg\min} \sum_{k=1}^{N} \left(\boldsymbol{\delta}[k] - \sum_{i=I_z[1]}^{I_z[p]} S_i \hat{\boldsymbol{z}}_i[k] - \sum_{j=I_u[1]}^{I_u[q]} \mathrm{t}_j \hat{\boldsymbol{u}}_j[k] \right)^2. \tag{G.8}$$

In the absence of noise (i.e. $\mathrm{t}_j = 0 \forall j$), each signal coefficient equals the projection $S_i = \boldsymbol{\delta}^T \hat{\boldsymbol{z}}_i$.

In principle, I_z and I_u can be recovered exactly if 1) all the signal and noise variance lie entirely within their respective indexed modes, and 2) those subspaces are non-overlapping.

For many problems, including this one, these conditions are unlikely to be met simultaneously. Therefore, trade-offs must be struck. The challenge is to retain as many signal modes as needed to approximate one's real data but not so many that crosstalk (i.e. overlaps) between $\hat{\boldsymbol{z}}_i$ and $\hat{\boldsymbol{u}}_j$ mixes information between the two subspaces.

We solve equation (G.8) by differentiating with respect to each of the S coefficients,

$$\frac{\partial}{\partial S_l} : \sum_{k=1}^{N} \left(\boldsymbol{\delta}[k] - \sum_{i=I_z[1]}^{I_z[p]} S_i \hat{\boldsymbol{z}}_i[k] - \sum_{j=I_u[1]}^{I_u[q]} \mathrm{t}_j \hat{\boldsymbol{u}}_j[k] \right) \cdot \hat{\boldsymbol{z}}_l[k] = 0. \tag{G.9}$$

Through orthonormality $\hat{\boldsymbol{z}}_l^T \hat{\boldsymbol{z}}_{i \neq l} = 0$, and therefore,

$$\sum_{k=1}^{N} \boldsymbol{\delta}[k] \hat{\boldsymbol{z}}_l[k] = S_l + \sum_{j=I_u[1]}^{I_u[q]} \mathrm{t}_j \sum_{k=1}^{N} \hat{\boldsymbol{u}}_j[k] \hat{\boldsymbol{z}}_l[k]. \tag{G.10}$$

If we let $\bar{\mathbf{Z}}$ and $\bar{\mathbf{U}}$ represent the truncated set of signal and noise eigenmodes respectively,

$$\bar{\mathbf{Z}} = \begin{bmatrix} | & & | \\ \hat{\boldsymbol{z}}_{I_z[1]} & \cdots & \hat{\boldsymbol{z}}_{I_z[p]} \\ | & & | \end{bmatrix}, \qquad \bar{\mathbf{U}} = \begin{bmatrix} | & & | \\ \hat{\boldsymbol{u}}_{I_u[1]} & \cdots & \hat{\boldsymbol{u}}_{I_u[q]} \\ | & & | \end{bmatrix}, \tag{G.11}$$

then this becomes a straightforward matrix equation,

$$\boldsymbol{S} = \bar{\mathbf{Z}}^T \boldsymbol{\delta} - \bar{\mathbf{Z}}^T \bar{\mathbf{U}} \mathfrak{t}. \tag{G.12}$$

A similar process can be used for the $\mathfrak{t}$ coefficients,

$$\frac{\partial}{\partial \mathfrak{t}_l} : \sum_{k=1}^{N} \left(\boldsymbol{\delta}[k] - \sum_{i=I_z[1]}^{I_z[p]} S_i \hat{\boldsymbol{z}}_i[k] - \sum_{j=I_u[1]}^{I_u[q]} \mathfrak{t}_j \hat{\boldsymbol{u}}_j[k] \right) \cdot [k] = 0, \tag{G.13}$$

$$\sum_{k=1}^{N} \boldsymbol{\delta}[k][k] = \mathfrak{t}_l + \sum_{i=I_z[1]}^{I_z[p]} S_j \sum_{k=1}^{N} \hat{\boldsymbol{z}}_i[k][k], \tag{G.14}$$

$$\mathfrak{t} = \bar{\mathbf{U}}^T \boldsymbol{\delta} - \bar{\mathbf{U}}^T \bar{\mathbf{Z}} \boldsymbol{S}. \tag{G.15}$$

Solving simultaneously reduces the system into an $Ax = b$ form,

$$\left(\mathbf{I} - \bar{\mathbf{U}}^T \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \bar{\mathbf{U}} \right) \boldsymbol{b} = \bar{\mathbf{U}}^T \boldsymbol{\delta} - \bar{\mathbf{U}}^T \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \boldsymbol{\delta}. \tag{G.16}$$

Note that if $\bar{\mathbf{Z}}$ is a complete basis, it is orthogonal and therefore $\bar{\mathbf{Z}}^T = \bar{\mathbf{Z}}^{-1}$. Consequently, the left-hand side of the equation becomes singular, $\left(\mathbf{I} - \bar{\mathbf{U}}^T \bar{\mathbf{U}} \right) = (\mathbf{I} - \mathbf{I})$.

The coefficients may be solved for through substitution or in a single step. If we define an overlap matrix $\mathbf{P} = \bar{\mathbf{U}}^T\bar{\mathbf{Z}}$ such that $P_{jl} = \hat{\boldsymbol{u}}_j^T\hat{\boldsymbol{z}}_l$, then

$$\bar{\mathbf{Z}}^T\boldsymbol{\delta} = \boldsymbol{S} + \mathbf{P}^T\mathbf{t},$$

$$\bar{\mathbf{U}}^T\boldsymbol{\delta} = \mathbf{t} + \mathbf{P}\boldsymbol{S},$$

and

$$\begin{bmatrix} \mathbf{I} & \mathbf{P}^T \\ \mathbf{P} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{S} \\ \mathbf{t} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{Z}}^T \\ \bar{\mathbf{U}}^T \end{bmatrix} \boldsymbol{\delta}. \tag{G.17}$$

The diagonal elements of this positive definite matrix equal 1, while the off-diagonal elements (which are projections) will have absolute values ≤ 1. More likely, the projections will be quite close zero. Consequently, we expect these matrices to be numerically stable and possess relatively low condition numbers.

The estimated signal and noise vectors are, respectively,

$$\hat{\boldsymbol{\delta}}_\kappa = \bar{\mathbf{Z}}\boldsymbol{S}, \quad {}_\eta = \bar{\mathbf{U}}\mathbf{t}. \tag{G.18}$$

On their own, $\hat{\boldsymbol{\delta}}_\kappa + \hat{\boldsymbol{\delta}}_\eta \neq \boldsymbol{\delta}$.

The two questions that arose after developing the theory of truncated expansion were 1) what combination of signal and noise modes should be used in the expansion, and 2) how

should we handle the fact that $\hat{\boldsymbol{\delta}}_\kappa + \hat{\boldsymbol{\delta}}_\eta \neq \boldsymbol{\delta}$? The first approach to answering question 1 utilized what we call *sequential truncation*. Sequential truncation takes $I_z = \{1, , p\}$ and $I_u = \{1, , q\}$. This method operates with the understanding that strong cross-talk between specific signal and noise modes might cause some combinations to be suboptimal.

We call the second approach *selective truncation*. Here expansion may occur over any combination of signal and noise modes. The process works by starting with values of p and q estimated through sequential truncation. With the noise modes fixed, each of the p modes is removed one at a time, leaving the others intact. Once the mode that minimizes one's chosen error metric is discovered, that mode is truncated and the process continues for the remaining $p-1$ modes until truncating signal no longer reduces the error. Once the optimal set of signal modes is determined, the same process takes place for the noise modes.

We found that truncation for zero-point noise was very sensitive to $\boldsymbol{\delta}$. When the algorithm reverses so that noise modes are truncated first, all noise modes would sometimes be eliminated, especially when σ_m was small. The results also appeared to depend on the "initial conditions" as communicated through p and q.

We answered the second question in three ways. The first estimated the signal without reintroducing the "missing" data, $\boldsymbol{\delta} - \left(\hat{\boldsymbol{\delta}}_\kappa + \hat{\boldsymbol{\delta}}_\eta\right)$. This exhibited exceptionally poor performance. The second reintroduced the missing data as pure signal,

$$\hat{\boldsymbol{\delta}}_\kappa \rightarrow \hat{\boldsymbol{\delta}}_\kappa + \left(\boldsymbol{\delta} - \left(\hat{\boldsymbol{\delta}}_\kappa + \hat{\boldsymbol{\delta}}_\eta\right)\right). \tag{G.19}$$

This produced substantially better results than ignoring the missing data, though not

enough to yield a genuine improvement over the status quo.

The third way expanded only the portion $\boldsymbol{\delta}'$ of $\boldsymbol{\delta}$ that lay in the intersection of the truncated subspaces,

$$\boldsymbol{\delta}'' = \bar{\mathbf{U}}\bar{\mathbf{U}}^T\bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\boldsymbol{\delta}. \tag{G.20}$$

In the extreme case where $\bar{\mathbf{U}}$ and $\bar{\mathbf{Z}}$ are non-overlapping, equation (G.20) returns nothing even though that would be an ideal condition for expansion. However, we know that the most important signal and noise modes correspond to large spatial structures. As such, they tend to overlap to a large enough degree that the intersection space is fairly large in practice.

We approximate the remainder $\boldsymbol{\delta}' \equiv \boldsymbol{\delta} - \boldsymbol{\delta}''$ as pure signal. This means

$$\begin{bmatrix} \mathbf{I} & \mathbf{P}^T \\ \mathbf{P} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{S} \\ \mathbf{t} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{Z}}^T \\ \bar{\mathbf{U}}^T \end{bmatrix} \boldsymbol{\delta}'', \tag{G.21}$$

which yields solutions of $\hat{\boldsymbol{\delta}}_\kappa = \bar{\mathbf{Z}}\boldsymbol{S} + \boldsymbol{\delta}'$ and $\hat{\boldsymbol{\delta}}_\eta = \bar{\mathbf{U}}\mathbf{t}$. The reintroduction of $\boldsymbol{\delta}'$ guarantees that $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}_\kappa + \hat{\boldsymbol{\delta}}_\eta$.

To assess the viability of this third option, we constructed performance matrices for $\sigma_m = \{0.02, 0.06, 0.2\}$. This PDF file displays the difference and signal-to-noise statistics for these three cases. Note, however, that these numbers are averaged over only 100 realizations and therefore suffer from reduced numerical precision. The columns list the number of noise modes retained q, while the rows list the number of signal modes retained

σ_m	m	p
0.02	1	1
0.06	2	15
0.20	$\sim$75	$\sim$35,000

Table G.1: Numbers of signal and noise modes in the truncated expansion that minimize the error metric for a fixed signal added to 100 vectors of random zero-point noise.

p. Retained modes include mode 1 through mode q or mode p. Table G.1 reports the best combinations to minimize $||\hat{\boldsymbol{\delta}}_\kappa - \boldsymbol{\delta}||_2$ for an earlier cell set of radius $8\,h^{-1}\mathrm{Mpc}$ with approximately 45,000 cells. The optimal combinations for signal-to-noise are similar.

To achieve optimum noise reduction, one should expand over more noise and more signal modes as σ_m increases. As the magnitude of the zero-point noise increases there will be a greater number of mode combinations that yield an improvement over the status quo. Though even at its best, the reduction in the norm of the noise is modest — between 0.3% and 5% with better performance occurring at larger σ_m.

It may be the case that a combination of signal and noise modes arrived at using techniques that differ from those presented above work best. The multitude of possibilities, though, limited our search for better combinations.

G.3 Gappy Reconstruction

Removal of low magnitude noise has presented one broad challenge — there is too much signal lying atop principle noise modes. In practice, our attempts to deproject data along noise modes carry away enough signal to be counterproductive. In this section, we

introduce a method that "zeroes out" data along the lowest order noise modes, then uses the signal eigenmodes to reintroduce the most likely values of the missing signal.

Let $\boldsymbol{\delta}_{\mathfrak{m}}$ represent a data vector $\boldsymbol{\delta}$ in noise-space. By virtue of being in noise-space, all of the information conveyed through the zero-points is organized into the first K dimensions. All zero-point noise can be eliminated by applying a mask function μ that equals zero along the first K dimensions, and 1 otherwise. This produces a *gappy* data vector whose k^{th} dimension equals

$$\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}[k] = \mu(k)\boldsymbol{\delta}_{\mathfrak{m}}[k]. \tag{G.22}$$

We refer to $\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}$ as a "gappy" data vector since the mask function introduces gaps (i.e. zeros) along several of its dimensions. The vector comprised of the nonzero elements of $\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}$ is known as the *support* of $\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}$, or $\mathrm{supp}\left(\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}\right)$.

Signal lies along every dimension, so masking out an element of $\boldsymbol{\delta}_{\mathfrak{m}}$ can potentially do more harm than good. It is crucial to only zero-out the n elements that are strictly necessary. In our problem, the zero-points have 1890 degrees of freedom, but approximately 90% of their variance lies along the first couple hundred dimensions. Deciding the proper mask function μ is therefore a critical part of this problem.

While equation (G.22) reflects the "gappy" part of this method, the "reconstruction" part combines the remaining data with one's signal model to effectively "fill in the gaps." This sort of technique in which an algorithm "guesses" the true signal within those gaps has been applied in fields like image processing to restore degraded data (see e.g. Everson

& Sirovich, 1995).

We expand the signal over its first $d \leq N - n$ signal eigenmodes *in noise-space*, or $\hat{\boldsymbol{v}}_i = \mathbf{U}^T \hat{\boldsymbol{z}}_i$. The size of d is constrained since it is counterproductive to solve for more signal coefficients in the set $\{a\}$ than there are degrees of freedom in the gappy data. We solve for $\{a\}$ by minimizing the difference of the data and the expansion in a least squared sense,

$$\underset{\{a\}}{\arg\min} \sum_k \left(\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}[k] - \sum_{i=1}^{d} a_i \hat{\boldsymbol{v}}_i[k] \right)^2_{\mathrm{supp}\left(\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}\right)} \tag{G.23}$$

The subscript $\mathrm{supp}\left(\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}\right)$ indicates that we only minimize over the support of the gappy data. The $N - d$ dimensions outside the support do not affect our solution and can be disregarded. For clarity, we notate the gappy data vector and signal eigenvectors for which the zeroed-out dimensions have been truncated as $\tilde{\boldsymbol{\delta}}_{\mathfrak{m}} \equiv \mathrm{supp}\left(\boldsymbol{\delta}_{\mathfrak{m}}^{(g)}\right)$ and $\tilde{\boldsymbol{v}}$, respectively. By setting the derivative of a_j to 0,

$$\frac{\partial}{\partial a_j} : \sum_k \left(\tilde{\boldsymbol{\delta}}_{\mathfrak{m}}[k] - \sum_{i=1}^{d} a_i \tilde{\boldsymbol{v}}_i[k] \right) \tilde{\boldsymbol{v}}_j[k] = 0, \tag{G.24}$$

we find

$$\tilde{\boldsymbol{\delta}}_{\mathfrak{m}}^T \tilde{\boldsymbol{v}}_j - \sum_{i=1}^{d} a_i \left(\tilde{\boldsymbol{v}}_i^T \tilde{\boldsymbol{v}}_j \right) = 0. \tag{G.25}$$

If we define $\tilde{\mathbf{V}}$ to be the $(N - n) \times d$ matrix

$$\tilde{\mathbf{V}} = \begin{bmatrix} | & & | \\ \tilde{\boldsymbol{v}}_1 & \cdots & \tilde{\boldsymbol{v}}_d \\ | & & | \end{bmatrix}, \tag{G.26}$$

and let $\mathbf{P} \equiv \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}$, then the problem reduces to a linear form,

$$\mathbf{P}\boldsymbol{a} = \tilde{\mathbf{V}}^T\tilde{\boldsymbol{\delta}}_{\mathfrak{m}}, \tag{G.27}$$

where $\boldsymbol{a}$ is a d-dimensional vector of expansion coefficients.

The reconstructed signal vector in noise-space, $\hat{\boldsymbol{\delta}}_{\mathfrak{s}}$, can be solved for by taking the product of $\boldsymbol{a}$ with the first d columns of $\mathbf{U}^T\mathbf{Z}$,

$$\hat{\boldsymbol{\delta}}_{\mathfrak{s}} = \mathbf{U}^T\mathbf{Z}\left[\mathbf{I}_d|\mathbf{0}\right]^T\boldsymbol{a}. \tag{G.28}$$

If one solves for $\boldsymbol{a}$ using multiple mask functions μ, reevaluation of $\mathbf{P}$ each time can become expensive. It is more efficient to perturb the identity matrix,

$$\mathbf{P} = \left[\mathbf{I}_d|\mathbf{0}\right]\left(\mathbf{I} - \tilde{\mathbf{V}}'\right)\left[\mathbf{I}_d|\mathbf{0}\right]^T, \tag{G.29}$$

where

$$\tilde{\mathbf{V}}'[i,j] \equiv \sum_{k=1}^{N} \hat{\boldsymbol{v}}_i[k]\hat{\boldsymbol{v}}_j[k] \;\; \forall k : \mu(k) = 0. \tag{G.30}$$

Likewise, because $\tilde{\mathbf{V}}$ is invoked during each solution it pays to precompute $\mathbf{U}^T\mathbf{Z}$, then eliminate the final $N - d$ columns and remove the k^{th} row $\forall k : \mu(k) = 0$.

Gappy reconstruction was originally tested using two amplified zero-point magnitudes $\sigma_m = \{0.02, 0.06\}$. In several ways, the results were similar to truncated reconstruction. Filling in multiple gaps simultaneously showed worse performance than filling in gaps one at a time (up to a point). The method was more effective as σ_m increased. There seemed to be a limiting σ_m below which gappy reconstruction was ineffective. Furthermore, the optimal d decreased with n. The range of amelioratory d values also decreased towards 1 with n. Finally, the results were sensitive to the number of realizations averaged over. We found that at least 50,000 realizations were necessary to reach stable averages.

Gappy reconstruction differed from truncated expansion is several important ways. For $\sigma_m = 0.02$, the optimal rms correction (defined to be $||\hat{\boldsymbol{\delta}}_{\mathfrak{s}} - \boldsymbol{\delta}_{\mathfrak{m}}||_2$ over the root of the number of realizations) occurred at $n = 1$ and $d = 18$ where $rms = 1.87$ (benchmark of 3.47). For $\sigma_m = 0.06$, the optimal rms correction occurred at $n = 1$ and $d = 660$ where $rms = 1.68$ (benchmark of 1.05). For a given n, the rms did not increase or decrease monotonically. There is a general concavity to the rms response, but with some small internal variance as well. There is an inflection point $n \approx 75$ (at least with $\sigma_m = 0.02$) where the rms starts to decrease with n. The magnitude of this inflection was not enough to catch the decreasing benchmarks, however.

Because replacing pixels (i.e. vector elements) one at a time displayed better performance than cleansing a handful at once, we altered the method to prioritize this strategy.

Pixels were "zeroed-out" one at a time, and multiple realizations were processed to determine the optimal number of signal modes d_{opt} over which to expand. We set d_{opt} to equal the number of signal modes that minimized the average difference between the signal vector $\boldsymbol{\delta}_{\mathfrak{m}}$ and the reconstructed data vector $\hat{\boldsymbol{\delta}}_{\mathfrak{s}}$,

$$d_{opt} = \arg\min_{d} \langle ||\hat{\boldsymbol{\delta}}_{\mathfrak{s}} - \boldsymbol{\delta}_{\mathfrak{m}}||_2 \rangle. \tag{G.31}$$

The solution in equation (G.28) remains the same save one exception. Because the reconstructed data vector is a linear combination of signal modes with a new vector of coefficients $\boldsymbol{a}$, the magnitude of each non-gappy pixel will change. By construction, the algorithm minimizes these differences such that in practice, if $d \approx N$ the discrepancies are potentially negligible. When d is small, however, the non-gappy pixels values can change appreciably. Since the noise component of $\boldsymbol{\delta}_{\mathfrak{m}}[k]$ equals zero for all $k > K$, allowing higher dimensional pixels to be modified at all is counterproductive. As a result, we modify the algorithm to restore all non-gappy pixel elements to their original values after reconstruction. If p is the index of the pixel being reconstructed, this means

$$\hat{\boldsymbol{\delta}}_{\mathfrak{s}}[k] = \boldsymbol{\delta}_{\mathfrak{m}}[k] \;\; \forall k \neq p. \tag{G.32}$$

Figure G.1 offers an example of gappy reconstruction for a single dimension after non-gappy pixel restoration is enabled. The response to signal reconstruction tends to follow curves of this nature with characteristic minimums, but which follow no particular func-

tional form. The caption of Figure G.1 explains the signal estimation process. In practice, the signal vector $\boldsymbol{\delta}_{\mathfrak{s}}$ used to seed the simulation should equal the true data vector when the noise component is small. In this way, variations in d_{opt} solutions due to changes in $\boldsymbol{\delta}_{\mathfrak{s}}$ can be mitigated.

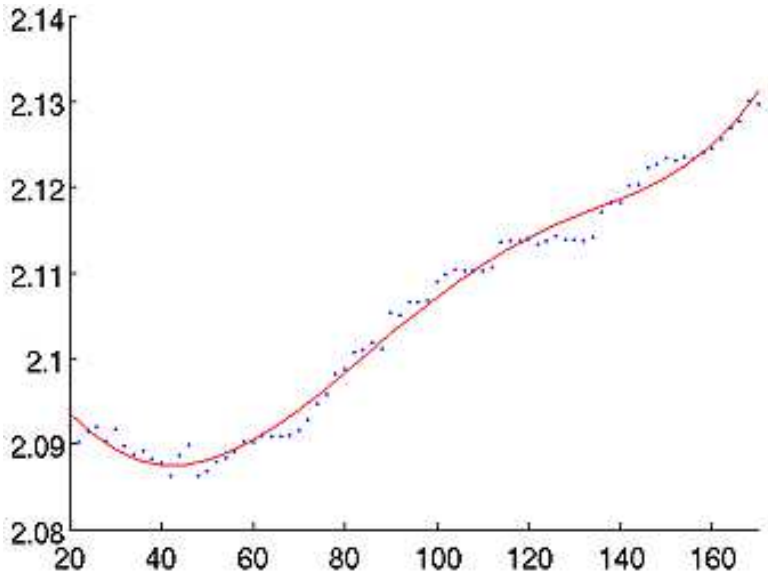

Figure G.1: Example of solving for d_{opt}. Random R7 data vectors $\boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)}$ with $\sigma_m = 0.15$ have their third pixel zeroed-out and their signal component $\boldsymbol{\delta}_{\mathfrak{s}}$ estimated using the method of gappy reconstruction. The number of signal modes d used in the expansion are reported on the horizontal axis. The average two-norm deviation between the original signal vector and reconstructed signal vector, $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2\rangle$, is reported on the vertical axis. Data realizations $\boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)} = \boldsymbol{\delta}_{\mathfrak{s}} + \boldsymbol{\delta}_{\mathfrak{t}}^{(\tau)}$ are assembled by adding 100,000 realizations of random noise $\boldsymbol{\delta}_{\mathfrak{t}}^{(\tau)}$ to the fixed signal. The red line is the best fit 5^{th} order polynomial. It is used to automatically quantify the location of d_{opt}. Here, $d_{opt} = 41$. If the resulting two-norm deviation of 2.088 is less than the benchmark $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)}||_2\rangle$ without reconstruction, then the correction is sustained. If the resulting two-norm deviation is larger, the pixel remains as is (i.e. with noise), and the process repeats for the fourth pixel, and so on.

Ultimately, we describe the *optimal gappy filter* with a collection of commands one can use to minimize $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2\rangle$. The commands are:

1. Zero out the i^{th} pixel if gappy reconstruction has been shown to be beneficial there,

2. Estimate the signal in pixel i using an expansion of $d_{opt}(i)$ signal modes,

3. Proceed to pixel $i + 1$ and repeat.

To investigate the quality of the optimal gappy filter we constructed a test case in which hundreds of thousands of realizations of zero-point noise with $\sigma_m = 0.15$ were added to a simulated $\boldsymbol{\delta}_{\mathfrak{s}}$. For each pixel, we measured $\langle ||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2 \rangle$ for a range of d values. We determined the d_{opt} that minimized the averaged 2-norm through visual inspection of curves the likes of which were presented in Figure G.1. A reporting of those measures is presented in Table G.2. We found that of the first 68 principle noise dimensions, only 12 benefited from reconstruction. Gappy reconstruction yielded an improvement in the 2-norm of only 1.7%.

pixel	d_{opt}	$\langle ||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2 \rangle$	
0	0	22.21351	0
1	33	22.09442	1
2	40	22.01892	1
3	66	21.99703	1
4	342	21.96861	1
5	58	21.98245	0
6	45	21.98623	0
7	1046	21.89354	1
8	40	21.8955	0
9	145	21.92666	0
10	1	22.03323	0
11	239	21.86167	1
12	391	21.86985	0
13	172	21.88834	0
14	5	21.94004	0
15	8	21.93819	0
16	230	21.87652	0
17	185	21.85644	1
18	28	21.94809	0

Table G.2 . . . continued

pixel	d_{opt}	$\langle \|\|\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}\|\|_2 \rangle$	
19	269	21.85527	1
20	28	21.92799	0
21	160	21.89929	0
22	599	21.86997	0
23	153	21.86464	0
24	233	21.88295	0
25	142	21.85645	0
26	186	21.86188	0
27	92	21.88375	0
28	232	21.90381	0
29	888	21.88283	0
30	345	21.87979	0
31	188	21.85314	1
32	153	21.85581	0
33	45	21.90432	0
34	51	21.95116	0
35	295	21.8628	0
36	275	21.90323	0
37	496	21.86012	0
38	254	21.91104	0
39	10	21.96007	0
40	211	21.92994	0
41	110	21.88431	0
42	75	21.8365	1
43	186	21.89609	0
44	322	21.84206	0
45	35	21.91218	0
46	495	21.84475	0
47	115	21.88174	0
48	172	21.90056	0
49	629	21.91529	0
50	313	21.87438	0
51	249	21.89121	0
52	429	21.88162	0
53	485	21.85615	0
54	363	21.85011	0
55	73	21.87892	0
56	491	21.84774	0
57	539	21.84837	0
58	278	21.93968	0

Table G.2 . . . continued

pixel	d_{opt}	$\langle\|\|\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}\|\|_2\rangle$	
59	536	21.83248	1
60	350	21.88772	0
61	108	21.91909	0
62	611	21.83784	0
63	566	21.83191	1
64	407	21.86003	0
65	989	21.84012	0
66	465	21.86525	0
67	394	21.85587	0
68	374	21.85254	0

Table G.2: Results of simulations used to construct the optimal gappy filter for fixed, random signal vector $\boldsymbol{\delta}_{\mathfrak{s}}$. Each row corresponds to a test over a single pixel. Pixels are ordered by the variance along the principle zero-point noise modes $\lambda_i^{(\eta)}$ from largest to smallest. For each pixel i, we add 200,000 realizations of zero-point noise $\boldsymbol{\delta}_{\mathfrak{t}}^{(\tau)}$ to $\boldsymbol{\delta}_{\mathfrak{s}}$ to generate $\boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)} = \boldsymbol{\delta}_{\mathfrak{s}} + \boldsymbol{\delta}_{\mathfrak{t}}^{(\tau)}$. We zero-out the i^{th} pixel and solve for the reconstructed signal $\hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}$ by expanding over a range of d signal modes. The values of d (second column) that minimize $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2\rangle$ (third column) are reported. The original benchmark difference between the signal and data, $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)}||_2\rangle = 22.21351$. If $\langle||\boldsymbol{\delta}_{\mathfrak{s}} - \hat{\boldsymbol{\delta}}_{\mathfrak{s}}^{(\tau)}||_2\rangle < \langle||\boldsymbol{\delta}_{\mathfrak{s}} - \boldsymbol{\delta}_{\mathfrak{m}}^{(\tau)}||_2\rangle$ for the i^{th} pixel, the value in the i^{th} pixel is marked for reconstruction and the fourth column is set to 1. If gappy reconstruction does not reduce the noise, the fourth column is set to zero.

After additional testing with various levels of zero-point noise, we reached several conclusions. First, d_{opt} decreases as σ_m increases. This is similar to truncated expansion. Second, gappy reconstruction failed to have a beneficial effect unless $\sigma_m > 0.1$, which potentially limits the usefulness of this method. In terms of developing the optimal gappy filter, the number of realizations was more important than the number of d_i points used for the either the polynomial fit or visual inspection, provided there was a large enough range of points. For example, it is better to increase the spacing between d values by a factor of 2 than it is to lower the number of realizations by a factor of 2.

APPENDIX G. APPENDIX G – ALTERNATIVE METHODS

One important difference between our analysis and the image reconstruction example provided in Everson & Sirovich (1995), is that their images contained $N \sim 10^4$ pixels, but their expansion occurred only over the first 50 signal eigenmodes. They demonstrated a reasonable signal recovery when 90% of their data pixels were zeroed-out. Even with this extreme mask function, there were 20 times more data pixels present than modes d expanded over. It is possible that since galaxy clustering signal possesses significant variance over all $N \sim 10^5$ dimensions, our expansion over $d_{opt} \sim 10^2$ signal eigenmodes is insufficient for adequate signal estimation.

Appendix H

Appendix H – Efficient Matrix Inversion for Noise of Limited Rank

Equation (7.6) requires the computationally expensive evaluation of $\mathbf{\Sigma}_{\kappa\eta}^{-1}(\sigma_m)$. One can modify the magnitude of the zero-point noise $\mathbf{\Sigma}_{\kappa\eta}^{-1} = (\mathbf{\Sigma}_\kappa + \mathbf{\Sigma}_\eta)^{-1}$ relatively quickly if $\mathbf{\Sigma}_\kappa^{-1}$ and the diagonalization $\mathbf{\Sigma}_\eta = \mathbf{U}\mathbf{\Lambda}^{(\eta)}\mathbf{U}^T$ are already known. Through expansion,

$$\begin{aligned}\mathbf{\Sigma}_{\kappa\eta}^{-1} &= (\mathbf{\Sigma}_\kappa + \mathbf{\Sigma}_\eta)^{-1} \\ &= \left(\mathbf{U}\left(\mathbf{U}^T\mathbf{\Sigma}_\kappa\mathbf{U}\right)\mathbf{U}^T + \mathbf{U}\mathbf{\Lambda}^{(\eta)}\mathbf{U}^T\right)^{-1} \\ &= \mathbf{U}\left(\mathbf{U}^T\mathbf{\Sigma}_\kappa\mathbf{U} + \mathbf{\Lambda}^{(\eta)}\right)^{-1}\mathbf{U}^T. \end{aligned} \tag{H.1}$$

By the Sherman-Morrison-Woodbury formula,

APPENDIX H. APPENDIX H – EFFICIENT MATRIX INVERSION FOR NOISE OF LIMITED RANK

$$(\mathbf{S}+\mathbf{\Lambda})^{-1} = \mathbf{S}^{-1} - \mathbf{S}^{-1}\left(\mathbf{S}^{-1}+\mathbf{\Lambda}^{-1}\right)^{-1}\mathbf{S}^{-1}. \tag{H.2}$$

Let $\mathbf{S} \equiv \mathbf{U}^T\mathbf{\Sigma}_\kappa\mathbf{U}$, and $\mathbf{T} \equiv \mathbf{S}^{-1} = \mathbf{U}^T\mathbf{\Sigma}_\kappa^{-1}\mathbf{U}$, and

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}, \qquad \mathbf{T}_c \equiv \begin{bmatrix} \mathbf{T}_{11} \\ \\ \mathbf{T}_{21} \end{bmatrix}, \qquad \mathbf{T}_r \equiv \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \end{bmatrix}, \tag{H.3}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{H.4}$$

where $\mathbf{T}_{11}$, $\mathbf{D} \in \mathbb{R}^{m\times m}$, $\mathbf{T}_{12} \in \mathbb{R}^{m\times(n-m)}$, and $\mathbf{T}_{22} \in \mathbb{R}^{(n-m)\times(n-m)}$. Then,

$$\begin{aligned} (\mathbf{S}+\mathbf{\Lambda})^{-1} &= \mathbf{T} - \mathbf{T}\left(\mathbf{T}+\mathbf{\Lambda}^{-1}\right)^{-1}\mathbf{T} \\ &= \mathbf{T} - \mathbf{T}\left(\mathbf{T}+\mathbf{\Lambda}^{-1}\right)^{-1}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{T} \\ &= \mathbf{T} - \mathbf{T}(\mathbf{\Lambda}\mathbf{T}+\mathbf{I})^{-1}\mathbf{\Lambda}\mathbf{T}. \end{aligned} \tag{H.5}$$

Simplify $(\mathbf{\Lambda}\mathbf{T}+\mathbf{I})^{-1}$ by partitioning it, then use the formula for the inverse of a partitioned matrix,

$$(\mathbf{\Lambda T}+\mathbf{I})^{-1} = \begin{bmatrix} \mathbf{I}+\mathbf{DT}_{11} & \mathbf{DT}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} \tag{H.6}$$
$$= \begin{bmatrix} (\mathbf{I}+\mathbf{DT}_{11})^{-1} & -(\mathbf{I}+\mathbf{DT}_{11})^{-1}\mathbf{DT}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Using equation (H.6), equation (H.5) simplifies to $(\mathbf{S}+\mathbf{\Lambda})^{-1} = \mathbf{T} - \mathbf{T}_c(\mathbf{I}+\mathbf{DT}_{11})^{-1}\mathbf{DT}_r$. If the eigenvalues stored in $\mathbf{\Lambda}^{(\eta)}$ have been calculated using $\sigma'_m = 1$, then equation (H.7) adjusts for an arbitrary scaling σ_m of the photometric zero-points,

$$\left(\mathbf{U}^T\mathbf{\Sigma}_\kappa\mathbf{U}+\mathbf{\Lambda}^{(\eta)}\right)^{-1} = \mathbf{T} - \sigma_m^2\mathbf{T}_c\left(\mathbf{I}+\sigma_m^2\mathbf{DT}_{11}\right)^{-1}\mathbf{DT}_r. \tag{H.7}$$

The reduced dimensionality of $\mathbf{D}$ and $\mathbf{T}_{11}$ renders an otherwise computationally expensive $n \times n$ matrix inverse into set of considerably smaller matrix products.

Bibliography

Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, AJ, 126, 2081

—. 2004, AJ, 128, 502

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297

Akrami, Y., Fantaye, Y., Shafieloo, A., et al. 2014, ApJL, 784, L42

Alcock, C., & Paczynski, B. 1979, , 281, 358

Aparicio Villegas, T., Alfaro, E. J., Cabrera-Caño, J., et al. 2010, AJ, 139, 1242

Astier, P., El Hage, P., Guy, J., et al. 2013, Astronomy and Astrophysics, 557, A55

Ballinger, W. E., Peacock, J. A., & Heavens, A. F. 1996, MNRAS, 282, 877

Balogh, M. L., McGee, S. L., Mok, A., et al. 2014, MNRAS, 443, 2679

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

BIBLIOGRAPHY

Bardeen, J. M., Steinhardt, P. J., & Turner, M. S. 1983, , 28, 679

Baum, W. A. 1959, PASP, 71, 106

Baum, W. A. 1962, in IAU Symposium, Vol. 15, Problems of Extra-Galactic Research, ed. G. C. McVittie, 390

Benítez, N., Dupke, R., Moles, M., et al. 2015, in Highlights of Spanish Astrophysics VIII, ed. A. J. Cenarro, F. Figueras, C. Hernández-Monteagudo, J. Trujillo Bueno, & L. Valdivielso, 148–153

Berlind, A. A., Frieman, J., Weinberg, D. H., et al. 2006, ApJS, 167, 1

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, , 367, 1

Blake, C., Collister, A., Bridle, S., & Lahav, O. 2007, MNRAS, 374, 1527

Blake, C., & Glazebrook, K. 2003, ApJ, 594, 665

Blake, C., Davis, T., Poole, G. B., et al. 2011, MNRAS, 415, 2892

Blanton, M. R., Lin, H., Lupton, R. H., et al. 2003, AJ, 125, 2276

Blanton, M. R., Dalcanton, J., Eisenstein, D., et al. 2001, AJ, 121, 2358

Bolstad, W. M. 2012, Understanding Computational Bayesian Statistics (Hoboken: John Wiley & Sons)

Bouchet, F. R., Colombi, S., Hivon, E., & Juszkiewicz, R. 1995, Astronomy and Astrophysics, 296, 575

BIBLIOGRAPHY

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Buchert, T. 1992, MNRAS, 254, 729

Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. 2000, AJ, 120, 1588

Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511

Carlson, J., Reid, B., & White, M. 2013, MNRAS, 429, 1674

Carrasco Kind, M., & Brunner, R. J. 2014, MNRAS, 438, 3409

Chung, D. J., Shiu, G., & Trodden, M. 2003, , 68, 063501

Coe, D., Benítez, N., Sánchez, S. F., et al. 2006, AJ, 132, 926

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505

Colless, M., Peterson, B. A., Jackson, C., et al. 2003, ArXiv Astrophysics e-prints, astro-ph/0306581

Collister, A. A., & Lahav, O. 2004, PASP, 116, 345

Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, AJ, 110, 2655

Cristóbal-Hornillos, D., Aguerri, J. A. L., Moles, M., et al. 2009, ApJ, 696, 1554

Csabai, I., Connolly, A. J., Szalay, A. S., & Budavári, T. 2000, AJ, 119, 69

BIBLIOGRAPHY

Csabai, I., Budavári, T., Connolly, A. J., et al. 2003, AJ, 125, 580

Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., & Lin, H. 2009, MNRAS, 396, 2379

Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003, 2MASS All Sky Catalog of point sources.

D'Abrusco, R., Staiano, A., Longo, G., et al. 2007, ApJ, 663, 752

Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93

Dalal, N., Doré, O., Huterer, D., & Shirokov, A. 2008, , 77, 123514

Davis, M., Meiksin, A., Strauss, M. A., da Costa, L. N., & Yahil, A. 1988, ApJL, 333, L9

Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465

Desai, S., Armstrong, R., Mohr, J. J., et al. 2012, ApJ, 757, 83

Doroshkevich, A. G., Zel'dovich, Y. B., & Syunyaev, R. A. 1978, , 22, 523

Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, MNRAS, 401, 1429

Efstathiou, G., Ellis, R. S., & Peterson, B. A. 1988, MNRAS, 232, 431

Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, AJ, 122, 2267

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560

Ellis, R. S., Colless, M., Broadhurst, T., Heyl, J., & Glazebrook, K. 1996, MNRAS, 280, 235

BIBLIOGRAPHY

Eriksen, H. K., Banday, A. J., Górski, K. M., Hansen, F. K., & Lilje, P. B. 2007, ApJL, 660, L81

Everson, R., & Sirovich, L. 1995, Journal of the Optical Society of America A, 12, 1657

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23

Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Levine, B., & Yahata, N. 2002, MNRAS, 330, 889

Finkbeiner, A. 2010, A Grand and Bold Thing: An Extraordinary New Map of the Universe Ushering (Free Press)

Fisher, K. B. 1995, ApJ, 448, 494

Fukugita, M., Ichikawa, T., Gunn, J. E., et al. 1996, AJ, 111, 1748

George, M. R., Leauthaud, A., Bundy, K., et al. 2011, ApJ, 742, 125

Gray, J., Szalay, A. S., Thakar, A. R., et al. 2004, eprint arXiv:cs/0408031, cs/0408031

Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332

Guo, H., Zehavi, I., & Zheng, Z. 2012, ApJ, 756, 127

Guth, A. H. 1981, , 23, 347

Guth, A. H., & Pi, S.-Y. 1982, Physical Review Letters, 49, 1110

BIBLIOGRAPHY

Hamilton, A. J. S. 1988, ApJL, 331, L59

—. 1992, ApJL, 385, L5

Hamilton, A. J. S. 1998, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185

Hawking, S. W. 1982, Physics Letters B, 115, 295

Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, Astronomy and Astrophysics, 523, A31

Hogg, D. W., Baldry, I. K., Blanton, M. R., & Eisenstein, D. J. 2002, ArXiv Astrophysics e-prints, astro-ph/0210394

Hogg, D. W., Finkbeiner, D. P., Schlegel, D. J., & Gunn, J. E. 2001, AJ, 122, 2129

Hogg, D. W., Blanton, M. R., Brinchmann, J., et al. 2004, ApJL, 601, L29

Howlett, C., Ross, A. J., Samushia, L., Percival, W. J., & Manera, M. 2015, MNRAS, 449, 848

Hu, W., & Haiman, Z. 2003, , 68, 063004

Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236

Ivezić, Ž., Lupton, R. H., Schlegel, D., et al. 2004, Astronomische Nachrichten, 325, 583

Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, ArXiv e-prints, arXiv:0805.2366

Jackson, J. C. 1972, MNRAS, 156, 1P

BIBLIOGRAPHY

Jarosik, N., Bennett, C. L., Dunkley, J., et al. 2011, ApJS, 192, 14

Jeong, D., & Komatsu, E. 2009, ApJ, 691, 569

Kaiser, N. 1987, MNRAS, 227, 1

Kelly, P. L., von der Linden, A., Applegate, D. E., et al. 2014, MNRAS, 439, 28

Khochfar, S., Silk, J., Windhorst, R. A., & Ryan, Jr., R. E. 2007, ApJL, 668, L115

Kuhn, T. 1996, The Structure of Scientific Revolutions, ISSR library (University of Chicago Press)

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Lawrence, E., Heitmann, K., White, M., et al. 2010, ApJ, 713, 1322

Lilly, S. J., Tresse, L., Hammer, F., Crampton, D., & Le Fevre, O. 1995, ApJ, 455, 108

Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118

Linder, E. V. 2005, , 72, 043529

Liske, J. 2000, MNRAS, 319, 557

López-Cruz, O., Barkhouse, W. A., & Yee, H. K. C. 2004, ApJ, 614, 679

López-Sanjuan, C., Balcells, M., Pérez-González, P. G., et al. 2010, Astronomy and Astrophysics, 518, A20

BIBLIOGRAPHY

Lupton, R., Gunn, J. E., Ivezić, Z., Knapp, G. R., & Kent, S. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 238, Astronomical Data Analysis Software and Systems X, ed. F. R. Harnden, Jr., F. A. Primini, & H. E. Payne, 269

Matsubara, T. 2008, , 78, 083519

McDonald, P. 2006, , 74, 103512

Myers, A. D., White, M., & Ball, N. M. 2009, MNRAS, 399, 2279

Neyrinck, M. C., Szapudi, I., & Rimes, C. D. 2006, MNRAS, 370, L66

Neyrinck, M. C., Szapudi, I., & Szalay, A. S. 2009, ApJL, 698, L90

Norberg, P., Baugh, C. M., Hawkins, E., et al. 2001, MNRAS, 328, 64

Oyaizu, H., Lima, M., Cunha, C. E., et al. 2008, ApJ, 674, 768

Padmanabhan, N., Schlegel, D. J., Finkbeiner, D. P., et al. 2008, ApJ, 674, 1217

Park, C., Vogeley, M. S., Geller, M. J., & Huchra, J. P. 1994, ApJ, 431, 569

Peebles, P. J. E. 1973, ApJ, 185, 413

—. 1980, The large-scale structure of the universe

Peebles, P. J. E. 1993, Principles of Physical Cosmology (New Jersey: Princeton University Press)

Peebles, P. J. E., & Yu, J. T. 1970, ApJ, 162, 815

Percival, W. J., Nichol, R. C., Eisenstein, D. J., et al. 2007, ApJ, 657, 645

Petrosian, V. 1976, ApJL, 209, L1

Pier, J. R., Munn, J. A., Hindsley, R. B., et al. 2003, AJ, 125, 1559

Planck Collaboration. 2011, A&A, 536, A1

Raccanelli, A., Bertacca, D., Pietrobon, D., et al. 2013, MNRAS, 436, 89

Rimes, C. D., & Hamilton, A. J. S. 2005, MNRAS, 360, L82

Roberts, G. O., Gelman, A., & Gilks, W. 1997, Ann. Appl. Probab., 7, 1, 110

Ryan, Jr., R. E., Hathi, N. P., Cohen, S. H., et al. 2007, ApJ, 668, 839

Sandage, A., Tammann, G. A., & Yahil, A. 1979, ApJ, 232, 352

Sawicki, M., & Thompson, D. 2006, ApJ, 648, 299

Schechter, P. 1976, ApJ, 203, 297

Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103

Schlafly, E. F., Finkbeiner, D. P., Schlegel, D. J., et al. 2010, ApJ, 725, 1175

Schlafly, E. F., Finkbeiner, D. P., Jurić, M., et al. 2012, ApJ, 756, 158

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525

Scoccimarro, R. 2004, , 70, 083007

Seo, H.-J., & Eisenstein, D. J. 2003, ApJ, 598, 720

Simpson, F., & Peacock, J. A. 2010, , 81, 043512

SkyServer. 2008, SkyServer DR6, `http://cas.sdss.org/dr6/en/`, accessed: 2015-07-01

Smith, J. A., Tucker, D. L., Kent, S., et al. 2002, AJ, 123, 2121

Starobinsky, A. A. 1982, Physics Letters B, 117, 175

Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810

Stubbs, C. W., & Tonry, J. L. 2006, ApJ, 646, 1436

SubbaRao, M., Frieman, J., Bernardi, M., et al. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4847, Astronomical Data Analysis II, ed. J.-L. Starck & F. D. Murtagh, 452–460

Sunyaev, R. A., & Zeldovich, Y. B. 1970, , 7, 3

Swanson, M. E. C., Percival, W. J., & Lahav, O. 2010, MNRAS, 409, 1100

Szalay, A. S., Matsubara, T., & Landy, S. D. 1998, ApJL, 498, L1

Szapudi, I. 2004, ApJ, 614, 51

Szapudi, I., & Szalay, A. S. 1996, ApJ, 459, 504

BIBLIOGRAPHY

Tagliaferri, R., Longo, G., Andreon, S., et al. 2003, Lecture Notes in Computer Science, 2859, 226

Tauber, J. A., Mandolesi, N., Puget, J.-L., et al. 2010, Astronomy and Astrophysics, 520, A1

Tegmark, M., Strauss, M. A., Blanton, M. R., et al. 2004, , 69, 103501

Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, , 74, 123507

Thakar, A. R., Szalay, A., Fekete, G., & Gray, J. 2008, Computing in Science and Engineering, 10, 30

Tucker, D. L., Kent, S., Richmond, M. W., et al. 2006, Astronomische Nachrichten, 327, 821

Uomoto, A., Smee, S., Rockosi, C., et al. 1999, in Bulletin of the American Astronomical Society, Vol. 31, American Astronomical Society Meeting Abstracts, 1501

Vanzella, E., Cristiani, S., Fontana, A., et al. 2004, Astronomy and Astrophysics, 423, 761

Verde, L., & Peiris, H. 2008, , 7, 9

Verde, L., Heavens, A. F., Percival, W. J., et al. 2002, MNRAS, 335, 432

Visvanathan, N., & Sandage, A. 1977, ApJ, 216, 214

Vogeley, M. S., & Szalay, A. S. 1996, ApJ, 465, 34

von der Linden, A., Best, P. N., Kauffmann, G., & White, S. D. M. 2007, MNRAS, 379, 867

Wadadekar, Y. 2005, PASP, 117, 79

Wang, Y., Bahcall, N., & Turner, E. L. 1998, AJ, 116, 2081

Weinberg, D. H. 1992, MNRAS, 254, 315

Weymann, R., Storrie-Lombardi, L., Sawicki, M., & Brunner, R., eds. 1999, Astronomical Society of the Pacific Conference Series, Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies

Wittman, D. 2009, ApJL, 700, L174

Yip, C.-W., Szalay, A. S., Carliles, S., & Budavári, T. 2011, ApJ, 730, 54

Yoon, J. H., Schawinski, K., Sheen, Y.-K., Ree, C. H., & Yi, S. K. 2008, ApJS, 176, 414

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Yu, Y., Zhang, P., Lin, W., Cui, W., & Fry, J. N. 2011, , 84, 023523

Zehavi, I., Blanton, M. R., Frieman, J. A., et al. 2002, ApJ, 571, 172

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2005, ApJ, 630, 1

—. 2011, ApJ, 736, 59

Zhao, G.-B., Saito, S., Percival, W. J., et al. 2013, MNRAS, 436, 2038

Vita

Mike Specian was born in Somerville, NJ in May, 1981. He attended Alpha Public School in Alpha, NJ through eighth grade, and graduated as valedictorian of Phillipsburg High School in Phillipsburg, NJ in 1999. Mike received his B.A. from Boston University in 2003 with a joint concentration in physics and astronomy. He graduated with honors and distinction after completing an undergraduate thesis on dynamic mass estimators under the tutelage of Professor Tereasa Brainerd.

After a short stint in the private sector, Mike enrolled in graduate school at Johns Hopkins University where he studied computational, statistical noise reduction techniques with Dr. Alex Szalay. He was an EARA Fellow at the Max Planck Institute for Astrophysics in Garching, Germany; a Mirzayan Science and Technology Policy Fellow at the National Academy of Sciences in Washington, DC; a scientific consultant for the US Global Change Research Program; and for three years served as the graduate representative of the Henry A. Rowland Department of Physics & Astronomy, both on campus and through the Amer-

ican Physical Society during the Coalition for National Science Funding's Congressional Visits Days.

Mike has spent considerable effort honing his skills as a physics teacher. After being a teaching assistant for JHU's undergraduate physics courses, he served as a Princeton Review MCAT instructor, a Center for Talented Youth course instructor, and private tutor. Beyond cosmology, Mike is very interested in science policy — particularly as it relates to climate, energy, sustainability, funding, education, and the areas in between.